Global Sensors Networks[1]

GSN Team

September 29, 2014

[1]The work presented in this paper was supported (in part) by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant no. 5005-67322 and by the Líon project supported by Science Foundation Ireland under grant no. SFI/02/CE1/I131.

Abstract

With the price of wireless sensor technologies diminishing rapidly we can expect large numbers of autonomous sensor networks being deployed in the near future. These sensor networks will typically not remain isolated but the need of interconnecting them on the network level to enable integrated data processing will arise, thus realizing the vision of a global "Sensor Internet." This requires a flexible middleware layer which abstracts from the underlying, heterogeneous sensor network technologies and supports fast and simple deployment and addition of new platforms, facilitates efficient distributed query processing and combination of sensor data, provides support for sensor mobility, and enables the dynamic adaption of the system configuration during runtime with minimal (zero-programming) effort. This paper describes the Global Sensor Networks (GSN) middleware which addresses these goals. We present GSN's conceptual model, abstractions, and architecture, and demonstrate the efficiency of the implementation through experiments with typical high-load application profiles. The GSN implementation is available from https://github.com/LSIR/gsn/.

This documentation will not be kept updated

It is intended for understanding the fundamentals of GSN, but for the latest changes and additions to GSN, please refer to the wiki
https://github.com/LSIR/gsn/wiki

Contents

Related Publications

Different parts of the work presented in this chapter are published in the form of articles in international conferences and workshops. Parts of this chapter are also published in the form of internal technical reports.

- *Infrastructure for data processing in large-scale interconnected sensor networks*, Karl Aberer , Manfred Hauswirth , Ali Salehi. Mobile Data Management (MDM), Germany, 2007.
- *GSN, Quick and Simple Sensor Network Deployment*, Ali Salehi, Karl Aberer. European conference on Wireless Sensor Networks (EWSN), Netherlands, 2007.
- *Zero-programming Sensor Network Deployment*, Karl Aberer , Manfred Hauswirth , Ali Salehi. Next Generation Service Platforms for Future Mobile Systems (SPMS), Japan, 2007.
- *A middleware for fast and flexible sensor network deployment*, Karl Aberer , Manfred Hauswirth , Ali Salehi. Very Large Data Bases (VLDB) Seoul, Korea, 2006.
- *Middleware support for the "Internet of Things"*, Karl Aberer , Manfred Hauswirth , Ali Salehi. 5. GI/ITG KuVS Fachgesprch "Drahtlose Sensornetze", Universitt Stuttgart, 2006.
- *The Global Sensor Networks middleware for efficient and flexible deployment and interconnection of sensor networks*, Karl Aberer , Manfred Hauswirth , Ali Salehi. Technical Report, LSIR-2006-006.
- *Global Sensor Networks*, Karl Aberer , Manfred Hauswirth , Ali Salehi. Technical Report, LSIR-2006-001.

Chapter 1

Introduction

Until now, research in the sensor network domain has mainly focused on routing, data aggregation, and energy conservation inside a single sensor network while the integration of multiple sensor networks has only been studied to a limited extent. However, as the price of wireless sensors diminishes rapidly we can soon expect large numbers of autonomous sensor networks being deployed. These sensor networks will be managed by different organizations but the interconnection of their infrastructures along with data integration and distributed query processing will soon become an issue to fully exploit the potential of this "Sensor Internet." This requires platforms which enable the dynamic integration and management of sensor networks and the produced data streams.

The Global Sensor Networks (GSN) platform aims at providing a flexible middleware to accomplish these goals. GSN assumes the simple model shown in Figure 1.1. A sensor network internally may use arbitrary multi-hop, ad-hoc routing algorithms to deliver sensor readings to one or more sink node(s). A sink node is a node which is connected to a more powerful base computer which in turn runs the GSN middleware and may participate in a (large-scale) network of base computers, each running GSN and servicing one or more sensor networks.

We do not make any assumptions on the internals of a sensor network other than that the sink node is connected to the base computer via a software wrapper conforming to the GSN API. On top of this physical access layer GSN provides so-called *virtual sensors* which abstract from implementation details of access to sensor data and define the data stream processing to be performed. Local and remote virtual sensors, their data streams and the associated query processing can be combined in arbitrary ways and thus enable the user to build a data-oriented "Sensor Internet" consisting of sensor networks connected via GSN.

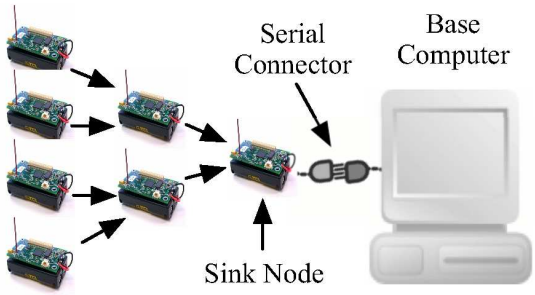

Figure 1.1: GSN model

1.1 Terminology

- **Global Sensor Networks** (*GSN*) defines both the project and the software described in this document.

- A **Wrapper** (*Wrapper*) is a piece of Java code that does the data acquisition for a specific type of device.

- A **Virtual Sensor** (*VS*) is the main component in *GSN*. It receives data from one or more *Wrapper*. It can combine their data, process and finally store it. A *VS* is defined in a single *VSD* and combines different pieces of software
  - One *VSP*
  - Zero or Many *Wrapper (s)*
- A **Virtual Sensor Description file** (*VSD*) is an XML file that contains the selection and the parametrization of the *VSP* and *Wrapper* that compose a *VS*. This file also contains the SQL statements that connect them together.

- A **Virtual Sensor Processing class** (*VSP*) is a piece of Java code that process and stores the data upon reception from the *Wrapper*.

1.2 Quick Start

GSN (for Global Sensor Networks) is a software project that started in 2005 at EPFL in the LSIR Lab by Ali Salehi, under the supervision of Prof. Karl Aberer. The initial goal was to provide a reusable software platform for the processing of data streams generated by wireless sensor networks. The project was successful, and was later reoriented towards a generic stream processing platform.

GSN acquires data, filters it with an intuitive, enriched SQL syntax, runs customisable algorithms on the results of the query, and outputs the generated data with its notification subsystem.

GSN can be configured to acquire data from various data sources. The high number of data sources in GSN allows for sophisticated data processing scenarios. In the unlikely event that your data sources are not supported, it is very easy to write a wrapper to make your hardware work with GSN (you can find more information about this in chapter 5).

GSN offers advanced data filtering functionalities through an enhanced SQL syntax. It is assumed that the reader has some knowledge of the Standard Query Language (SQL). Using it for basic operations is fairly intuitive and you should be able to start using it from the examples provided in this document.

Chapter 2

GSN Architecture

GSN uses a container-based architecture for hosting virtual sensors. Similar to application servers, GSN provides an environment in which sensor networks can easily and flexibly be specified and deployed by hiding most of the system complexity in the GSN Server. Using the declarative specifications, virtual sensors can be deployed and reconfigured in GSN Servers at runtime. Communication and processing among different GSN Servers is performed in a peer-to-peer style through standard Internet and Web Services protocols. By viewing GSN Servers as cooperating peers in a decentralized system, we tried avoid some of the intrinsic scalability problems of many other systems which rely on a centralized or hierarchical architecture. Targeting a "Sensor Internet" as the long-term goal we also need to take into account that such a system will consist of "Autonomous Sensor Systems" with a large degree of freedom and only limited possibilities of control, similarly as in the Internet.

Figure 2.1 shows the layered architecture of a GSN Server.

Each GSN server hosts a number of virtual sensors it is responsible for. The virtual sensor manager (VSM) is responsible for providing access to the virtual sensors, managing the delivery of sensor data, and providing the necessary administrative infrastructure. The VSM has two subcomponents: The life-cycle manager (LCM) provides and manages the resources provided to a virtual sensor and manages the interactions with a virtual sensor (sensor readings, etc.). The input stream manager (ISM) is responsible for managing the streams, allocating resources to them, and enabling resource sharing among them while its stream quality manager subcomponent (SQM) handles sensor disconnections, missing values, unexpected delays, etc., thus ensuring the QoS of streams. All data from/to the VSM passes through the storage layer which is in charge of providing and managing persistent storage for data streams. Query processing in turn relies on all of the above layers and is done by the query manager (QM) which includes the query processor being in charge of SQL parsing, query planning, and execution of queries (using an adaptive query execution plan). The query repository manages all registered queries (subscriptions) and defines and maintains the set of currently active queries for the query processor. The notification manager deals with the delivery of events and query results to registered, local or remote virtual sensors. The notification manager has an extensible architecture which allows the user to largely customize its functionality, for example, having results mailed or being notified via SMS.

The top three layers of the architecture deal with access to the GSN server. The interface layer provides access functions for other GSN servers and via the Web (through a browser or via web services). These functionalities are protected and shielded by the access control layer providing access only to entitled parties and the data integrity layer which provides data

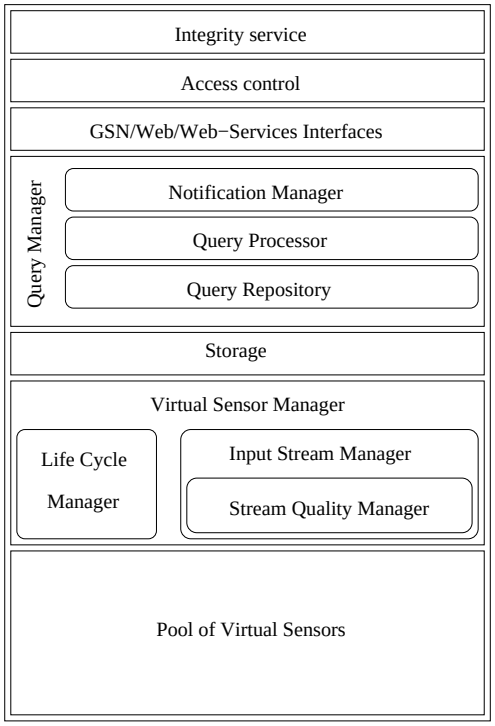

Figure 2.1: GSN Server architecture

integrity and confidentiality through electronic signatures and encryption. Data access and data integrity can be defined at different levels, for example, for the whole GSN server or at a virtual sensor level.

In connection with RFID tags this "plug-and-play" feature of GSN provides new and interesting types of mobility which we will investigate in future work. For example, an RFID tag may store queries which are executed as soon as the tag is detected by a reader, thus transforming RFID tags from simple means for identification and description into a GSN server for physically mobile queries which opens up new and interesting possibilities for mobile information systems.

2.1 Data Acquisition

Before filtering and processing data, *GSN* needs to receive it. *GSN* considers two types of data sources: event-based and polling-based. In the first case, data is sent by the source and a *GSN* method is called when it arrives. Serial ports, network (TCP or UDP) connections, wireless webcams fall in this case. In the latter one, GSN periodically asks the source for new data. This is the case of an RSS feed or a POP3 email account.

2.1.1 *GSN* Wrappers

GSN can receive data from various data sources. This is done by using so called wrappers. They are used to encapsulate the data received from the data source into the standard GSN data model, called a StreamElement. A StreamElement is an object representing a row of a SQL

table. Each wrapper is a Java class that extends the AbstractWrapper parent class. Usually a wrapper initializes a specialized third-party library in its constructor. It also provides a method which is called each time the library receives data from the monitored device. This method will extract the interesting data, optionally parse it, and create one or more StreamElement(s) with one or more columns. From this point on, the received data has been mapped to a SQL data structure with fields that have a name and a type. *GSN* is then able to filter this using its enhanced SQL-like syntax. You will learn more about that in section 2.7.3. A wrapper is implemented in a Java class. For simplicity, *GSN* uses short names to refer to these wrappers. These associations are defined in the file `conf/wrappers.properties`. For now on it is assumed that you use the default names provided at installation time.

2.2 Data Filtering and Processing

GSN provides two complementary mechanisms to work on data. The first one is based on a SQL syntax enhanced with specialized semantics for timed sliding windows and event counting. The second one allows to manipulate data with specialized programs called virtual sensors. GSN comes with a library of virtual sensors that you can use without programming.

GSN always processes the data according to a virtual sensor configuration. If you only want to use the SQL filtering mechanism, without any data trans- formation, you can use the BridgeVirtualSensor (see **??**). If you have more sophisticated needs, you can write your own virtual sensor processing class (See appendix B.1). If you don't want to use the SQL filtering mechanism, simply select all data from the wrapper.

2.2.1 Virtual Sensors

The key abstraction in GSN is the *virtual sensor*. Virtual sensors abstract from the implementation details of the data aource to sensor data and correspond either to a data stream received directly from sensors or to a data stream derived from other virtual sensors. A virtual sensor can be any kind of data producer, for example, a real sensor, a wireless camera, a desktop computer, or any combination of virtual sensors. A virtual sensor may have any number of input data streams and produces exactly one output data stream (with predefined format) based on the input data streams and arbitrary local processing. The specification of a virtual sensor provides all necessary information required for deploying and using it, including (1) metadata used for identification and discovery, (2) the details of the data streams which the virtual sensor consumes and produces (3) an SQL-based specification of the stream processing (filtering and integration) performed in a virtual sensor, (4) the processing class which performs the more advanced and complex data processing (if needed) on the output stream before releasing it and (5) functional properties related to persistency, error handling, life-cycle, management, and physical deployment.

To support rapid deployment, the virtual sensors are provided in human readable declarative forms (XML). Figure 2.1 shows an example which defines a virtual sensor that reads two temperature sensors and in case both of them have the same reading above a certain threshold in the last minute, the virtual sensor returns the latest picture from the webcam in the same room together with the measured temperature.

```
<virtual-sensor name="room-monitor" priority="10"
         protected="false" >
  <processing-class>
     <class-name>gsn.vsensor.BridgeVirtualSensor</class-name>
```

```
    <init-params/>
    <output-structure>
       <field name="image" type="binary:jpeg" />
       <field name="temp"  type="int" />
    </output-structure>
  </processing-class>
  <life-cycle pool-size="10" />
  <addressing>
      <predicate key="geographical">BC143</predicate>
      <predicate key="usage">room monitoring</predicate>
      <predicate key="latitude">46.5214</predicate>
      <predicate key="longitude">6.5676</predicate>
  </addressing>
  <storage history-size="10h" />
  <streams>
      <stream name="cam">
         <source name="cam"  storage-size="1" >
            <address wrapper="remote">
               <predicate key="geographical">BC143</predicate>
               <predicate key="type">Camera</predicate>
            </address>
            <query>select * from WRAPPER</query>
         </source>
         <source name="temperature1" storage-size="1m" >
            <address wrapper="remote">
               <predicate key="type">temperature</predicate>
               <predicate key="geographical">BC143-N</predicate>
            </address>
            <query>select AVG(temp1) as T1 from WRAPPER</query>
         </source>
         <source name="temperature2"  storage-size="1m" >
            <address wrapper="remote">
               <predicate key="type">temperature</predicate>
               <predicate key="geographical">BC143-S</predicate>
            </address>
            <query>select AVG(temp2) as T2 from WRAPPER</query>
         </source>
         <query>
            select cam.picture as image, temperature.T1 as temp
            from   cam, temperature1
            where  temperature1.T1 > 30 AND
                   temperature1.T1 = temperature2.T2
         </query>
      </stream>
   </streams>
</virtual-sensor>
```

Listing 2.1: A virtual sensor definition

A virtual sensor has a unique name (the `name` attribute in line 1) and can be equipped with a set of key-value pairs representing the logical addressing of the virtual sensor (lines 12–17), i.e., associated with metadata. The addressing information can be registered and discovered in GSN and other virtual sensors can use either the unique name or logical addressing based on the metadata to refer to a virtual sensor. We have defined certain addressing keys which are specifically used by the GSN's web interface. In GSN if a given virtual sensor has the addressing values for the both `latitude` (line 15) and `longitude` (line 16) keys, the default GSN web interface uses these geographical locations to show the sensor on the global map.

The example specification above defines a virtual sensor with three input streams which are identified by their metadata[1], i.e., by logical addressing. For example, the first temperature

[1]Note that the support for distributed directory/registry service had been removed from GSN's source code thus as of September 29, 2014, we only support physical addressing for identifying the data sources.

sensor is addressed by specifying two requirements on its metadata, namely that it is of type temperature sensor and at a certain physical certain location. By using multiple input streams Figure 2.1 also demonstrates GSN's ability to access multiple stream producers simultaneously. For the moment, we assume that the input streams (two temperature sensors and a webcam) have already been defined in other virtual sensor definitions (how this is done, will be described below).

In GSN, data streams are temporal sequences of timestamped tuples (also known as `Stream Elements`). This is in line with the model used in most stream processing systems. The structure of the output data stream a virtual sensor produces is encoded in XML as shown in lines 6 – 9 (the `output-structure` part). The structure of the input streams is learned from the respective specifications of their virtual sensor definitions.

In GSN data stream processing is separated into three stages:

- processing applied to sources (lines 26, 33, and 40).
- processing for combining data from the different input streams and producing the temporary output stream (lines 43-46).
- producing the final output stream by passing the temporary output stream from a processing class (a processing logic represented in some programming languages). This part is presented by lines 3 – 10. Note that as the final output is produced by the processing class, the actual output structure of the virtual sensor should strictly conform the output format of the processing class [2].

To specify the processing of the sources we use SQL queries which refer to the actual data source by the reserved keyword `WRAPPER` (the data sources are logically represented as relational tables all of which are called `wrapper`). The attribute `wrapper="remote"` indicates that the data stream is obtained through the network from another virtual sensor which can be located in any other GSN instance accessible through the network.

In the case of a directly connected local sensor, the `wrapper` attribute would reference the required wrapper[3]. For example, `wrapper="tinyos"` would denote a TinyOS-based sensor whose data stream is accessed via GSN's TinyOS wrapper [4]. GSN already includes wrappers for all major TinyOS platforms (Mica2, Mica2Dot, etc.), for wired and wireless (HTTP-based) cameras (e.g., AXIS 206W), several RFID readers (Texas Instruments, Alien Technology), Bluetooth devices, Shockfish, WiseNodes, epuck robots, etc. The implementation effort for wrappers is rather low, for example, the RFID reader wrapper has 50 lines of code (LOC), the TinyOS wrapper has 120 LOC, and the generic serial wrapper has 180 LOC.

In the given example the output stream joins the data received from two temperature sensors and returns a camera image if certain conditions on the temperature are satisfied (lines 43–46). To enable the SQL statement in lines 43–46 to produce the output stream, it needs to be able to reference the required sources which is accomplished by the `name` attribute (lines 21, 28, and 35) that defines a symbolic name for each stream source.

The definition of the structure of the output stream directly relates to the data stream processing that is performed by the virtual sensor's processing class and needs to

[2] As of September 29, 2014, the order and the types should be exactly match.

[3] As of September 29, 2014, all the wrappers have to written in Java language. The actual code for accessing the sensor can be written in any language as long as there is a possibility of communicating the data to the hardware through Java (e.g., interfacing Java to existing C code or the serial ports).

[4] In GSN, we have multiple TinyOS wrappers each corresponding to different versions and packet formats. Those details are out of the scope of this chapter.

be consistent with it. GSN provides multiple processing classes each of which are designed to perform different tasks (e.g., charts, network plots, filtering, ...). In our example we are using `gsn.vsensor.BridgeVirtualSensor` as our processing class. The `gsn.vsensor.BridgeVirtualSensor` class is special in the sense that unlike most of the other GSN's processing classes, this class does not perform any further processing on its input stream thus it does not alter the data nor the structure of its input.

Since the structure of the virtual sensor output is not altered through using the `gsn.vsensor.BridgeVirtualSensor` processing class hence the final structure of the virtual sensor's output is determined through the SQL statement at line 43, we need to make sure that, the data fields in the `select` clause matches the definition of the output structure in lines 6–9 (the order is important). It is recommended to use `gsn.vsensor.BridgeVirtualSensor` as long as the processing preformed in the virtual sensor through the SQL queries are sufficient enough and no further processing is required before publishing the sensor data to the outside.

In the design of GSN specifications we decided to separate the temporal aspects from the relational data processing using SQL. The temporal processing is controlled by various attributes provided in the input and output stream specifications, e.g., the attribute `storage-size` (lines 21, 28, and 35) defines the window size used for producing the input stream's data elements. Due to its specific importance the temporal processing will be discussed in detail in Section 2.4.

In addition to the specification of the data-related properties a virtual sensor also includes high-level specifications of functional properties: The `priority` attribute (line 1) controls the processing priority of a virtual sensor, the `<life-cycle>` element (line 11) enables the control and management of resources provided to a virtual sensor such as the maximum number of threads/queues available for processing, the `<storage>` element (line 18) allows the user to control how output stream data is persistently stored.

For example, in Figure 2.1 the `priority` attribute in line 1 assigns a priority of 10 to this virtual sensor (1 is the lowest priority and 20 the highest, default is 10), the `<life-cycle>` element in line 11 specifies a maximum number of 10 threads, which means that if the pool size is reached, data will be dropped (if no pool size is specified, it will be controlled by GSN depending on the current load), the `<storage>` element in line 18 defines that the output stream's data elements of the last 10 hours (`history-size` attribute) are stored to enable off-line processing. The `storage-size` attribute in line 21 defines the window size of 1 stream element. That's the most recent image taken by the webcam irrespective of the time it was taken.

In GSN, we can specify the set of values either by time or count. In the count based representation one only presents the values through integers. For instance `slide='2'` or `history-size='100'`. The count based representation is consists of an integer directly postfixed (without any space characters) with one of the time measurement units. As of September 29, 2014, we have `d`,`h`,`m`,`s` time measurement units which are corresponding to days, hours, minutes and seconds. As a time based example, we might have `storage-size='1m'`.

The `storage-size` attributes in lines 28 and 35 define a window of one minute for the amount of sensor readings subsequent queries will be run on, i.e., the `AVG` operations in lines 33 and 40 are executed on the sensor readings received in the last minute which of course depends on the rate at which the underlying temperature virtual sensor produces its readings. Note that when the `storage-size` is anything other than *1*, the virtual sensor author should be aware of the possibility of duplicated stream elements (discussed in more detail in section 2.4).

The query producing the output stream (lines 43–46) also demonstrates another interesting capability of GSN as it also mediates among three different flavors of queries: The virtual sensor itself uses continuous queries on the temperature data, a "normal" database query on the camera data and produces a result only if certain conditions are satisfied, i.e., a notification

analogous to pub/sub or active rules.

Virtual sensors are a powerful abstraction mechanism which enables the user to declaratively specify sensors and combinations of arbitrary complexity. Virtual sensors can be defined and deployed to a running GSN instance at any time without having to stop the system. Also dynamic unloading is supported but should be used carefully as unloading a virtual sensor may have undesired (cascading) effects.

2.3 Data publishing

2.3.1 Web Interface

GSN ships with an elegant and easy to use web interface. The only thing you have to do is to open a web browser and go the following address: http: //127.0.0.1:22001 .

2.3.1.1 GoogleMaps integration

GSN can associate your data with GPS positions and then display these on a world map retrieved from Google's GoogleMaps service. You need a special identification key from Google. For more information, please refer to the docu- mentation file doc/README.txt, section 'How to use GoogleMaps with GSN'.

2.4 Data stream processing and time model

Data stream processing has received substantial attention in the recent years in other application domains, such as network monitoring or telecommunications. As a result, a rich set of query languages and query processing approaches for data streams exist on which we can build. A central building block in data stream processing is the time model as it defines the temporal semantics of data and thus determines the design and implementation of a system. Currently, most stream processing systems use a global reference time as the basis for their temporal semantics because they were designed for centralized architectures in the first place. As GSN is targeted at enabling a distributed "Sensor Internet," imposing a specific temporal semantics seems inadequate and maintaining it might come at unacceptable cost. GSN provides the essential building blocks for dealing with time, but leaves temporal semantics largely to applications allowing them to express and satisfy their specific, largely varying requirements. In our opinion, this pragmatic approach is viable as it reflects the requirements and capabilities of sensor network processing.

In GSN a data stream is a set of timestamped tuples also known as Stream Elements. The order of the data stream is derived from the ordering of the timestamps and GSN provides basic support for managing and manipulating the timestamps. The following essential services are provided:

1. a local clock at each GSN Server
2. implicit management of a timestamp attribute (reserved field called `TIMED`)[5][6]

[5] All timestamps in GSN are represented in milliseconds using 64-bit integers.

[6] As the timestamp (e.g., the `TIMED` field) is always present, it is not required to specify the `TIMED` field in the `output-structure` section of the virtual sensors. In fact, specifying the `TIMED` field in the output structure causes error therefore GSN refuses to load the virtual sensor.

3. automatic timestamping of tuples upon arrival at the GSN in case the tuples (stream elements) don't have any timestamp (no `TIMED` field available)

4. a windowing mechanism which allows the user to define count- or time-based windows on data streams.

5. a sliding mechanism which allows the user to define count- or time-based sliding behaviors on data streams.

In this way it is always possible to trace the temporal history of data stream elements throughout the processing history. Multiple time attributes can be associated with data streams (as long as only one of them called `TIMED`) and can be manipulated through SQL queries. Thus sensor networks can be used as observation tools for the physical world, in which network and processing delays are inherent properties of the observation process which cannot be made transparent by abstraction. Let us illustrate this by a simple example: Assume a bank is being robbed and images of the crime scene taken by the security cameras are transmitted to the police. For the insurance company the time at which the images are taken in the bank will be relevant when processing a claim, whereas for the police report the time the images arrived at the police station will be relevant to justify the time of intervention. Depending on the context the robbery is thus taking place at different times.

As tuples (sensor readings) are timestamped, queries can also deal explicitly with time. For example, the query in lines 43–46 of Figure 2.1 could be extended such that it explicitly specifies the maximum time interval between the readings of the two temperatures and the maximum age of the readings. This would additionally require changes in the source definitions as the sources then must provide this information (more detailed example below), and also the averaging of the temperature readings (lines 33 and 40) would have to be changed to be explicit in respect to the time dimension.

In order to concretely show the time management inside GSN, we would like to simulate above scenario through two different virtual sensors (only the input stream parts presented). Say there exist a virtual sensor called *camera-vs* hosted on a GSN server which listens to port `80` on a machine with IP address of `1.2.3.4`. The virtual sensor used by the insurance and the one used by the police are depicted in figures 2.2 and 2.3. The stream specified in figure 2.2 has a query in line 7 for retrieving both the picture and the time stamp from the remote virtual sensor therefore the remote timestamp is used by GSN for the internal calculations. Now consider the stream specified in figure 2.3 which has a small change compared to the one in figure 2.2, the latter is not selecting the timestamp field hence GSN automatically adds the local reception time to every tuple it receives from the remote source.

In order to further elaborate the time management issue, consider the stream source specified in figure 2.4. This example combines both the local time and remote time in order to measure the latency associated with each tuple and uses the latency as a condition as the selection criteria (e.g., only accepting the tuples which are not delayed by the network for more than 5 milliseconds).

```
<source name="cam"  storage-size="1" >
   <address wrapper="remote">
      <predicate key="host">1.2.3.4</predicate>
      <predicate key="port">80</predicate>
      <predicate key="name">camera-vs</predicate>
   </address>
   <query>select PICTURE, TIMED from WRAPPER</query>
</source>
 <query>
   select PICTURE, TIMED from cam
</query>
```

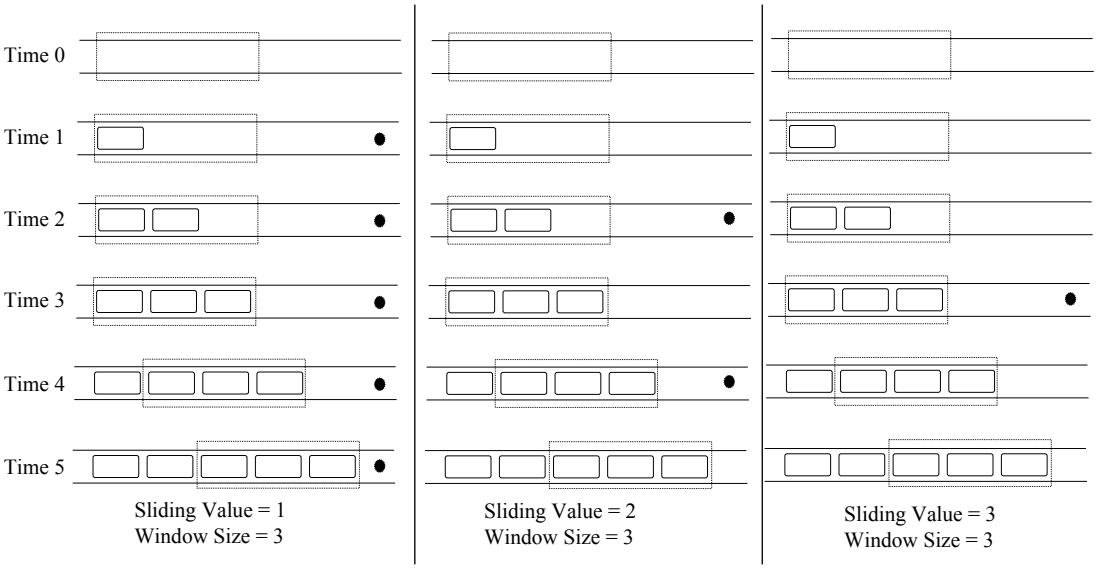

Figure 2.2: Illustration of the different sample sliding and window values.

```
</stream>
```

Listing 2.2: A stream using the remote timestamp.

```
<stream name="cam">
   <source name="cam"  storage_size="1" >
      <address wrapper="remote">
         <predicate key="host">1.2.3.4</predicate>
         <predicate key="port">80</predicate>
         <predicate key="name">camera_vs</predicate>
      </address>
      <query>select PICTURE from WRAPPER</query>
   </source>
   <query>
      select PICTURE, TIMED from cam
   </query>
</stream>
```

Listing 2.3: A stream using the local (arrival) timestamp.

```
<stream name="cam">
   <source name="cam"  storage_size="1" >
      <address wrapper="remote">
         <predicate key="host">1.2.3.4</predicate>
         <predicate key="port">80</predicate>
         <predicate key="name">camera_vs</predicate>
      </address>
      <query>select PICTURE, TIMED as REMOTE_TIMED from WRAPPER</query>
   </source>
   <query>
      select PICTURE, REMOTE_TIMED AS TIMED from cam where
      (cam.TIMED - cam.REMOTE_TIMED) < 5
   </query>
</stream>
```

Listing 2.4: A stream using both local and remote timestamps.

In order to deal with the streaming data, the standard way is to specify a query with at least two extra properties associated with it, window size and sliding value. The window size is used

to limit the actual data used for the processing (execution) to a certain range in time or number of values. The sliding value is introduced to specify the execution condition for the query. The execution of the query is triggered whenever the sliding condition is satisfied implying a possibly infinitely long periodic execution of the query, therefore in stream processing systems, continuous queries are executed whenever the sliding occurs.

For instance, one can express the interest of obtaining the average of a temperature sensor over the last 10 minutes, and doing so periodically every 2 minutes, by simply providing the window size of 10 minutes and sliding value of 2 minutes to the stream processing engine. As indicated before, each time the sliding condition is satisfied (e.g., 2 minutes passed from the previous execution) the actual action, computing the average over the last 10 minutes, is performed. Note that in some research papers the execution of the action is also called *movement of the sliding window*.

The temporal processing in GSN is defined using the sliding and window values. Every data source in GSN can have at most one `slide`[7] and `storage-size`[8] attributes. Both values can be represented in the form of count-based or time-based values (described earlier in this section). Figure 2.2 visually represents the query execution inside GSN with different sliding and window values. We used a black dot in the figure to represent the triggering of execution. For instance, if both the window size and the sliding values are 3, and say we have received 5 stream elements in total, our continuous query have been executed only once (at the *Time 3*) during its life time. One can extend above paradigm to create virtual sensors to support the integration of continuous and historical data. For example, if the user wants to be notified when the temperature is 10 degrees above the average temperature in the last 24 hours, he/she can simply define two sources, getting data from the same wrapper but with different window sizes, i.e., 1 (count) and 24h (time), and then simply write a query specifying the original condition with these sources.

The production of a new output stream element of a virtual sensor is always triggered by the arrival of a data stream element from one of its input streams, thus processing is event-driven. As described before, a stream can have multiple sources. Once the window of one of the sources of a stream slides, the following processing steps are performed:

1. Based on the timestamps for each stream the stream elements are selected according to the definition of the time window and the resulting sets of relations are unnested into flat relations.
2. The queries defined on the source are evaluated and stored into temporary relations.
3. The stream query for producing the input of the processing class is executed based on the temporary relations.
4. The resulted stream elements are forwarded to the processing class.
5. The output of the processing class is stored and simultaneously forwarded (notification) to all consumers of the virtual sensor.

Figure 2.3 shows the logical data flow inside a GSN node.

Additionally, GSN provides a number of attributes in the virtual sensor file to control data rates.

At the source level by providing the `sampling-rate` attribute to allow the dropping of stream elements with some random probability for load shedding. The values used for

[7]Default value is 1, therefore this attribute can be omitted
[8]No default value defined

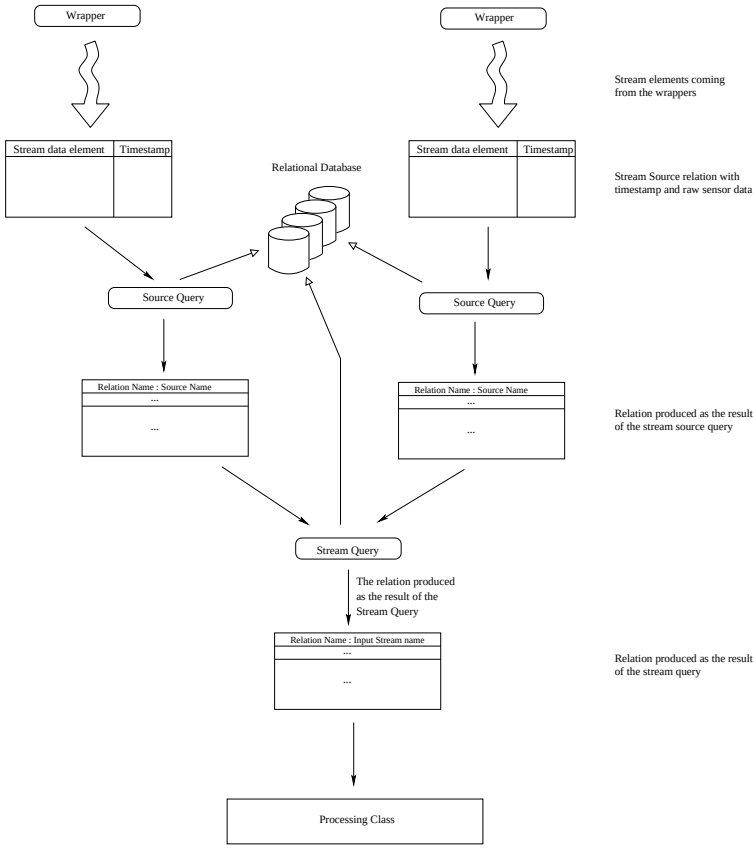

Figure 2.3: Conceptual data flow in a GSN node

`sampling-rate` are float numbers between 0 to 1. For instance, if one has a temperature source that keeps producing data with very high rate, one might want to sample the produced values thus making the processing lighter. For instance if one sets the sampling-rate to 0.75, any received stream element from the wrapper is going to be included in the window (the window and sliding values are explained above) with a probability of 75 out of 100. Thus, on average 25 random stream elements will be dropped out of the last 100 elements. In most of the cases one typically sets the rate control attributes to "1" to make sure nothing is dropped.

Two other rate controls that function in a different manner are:

- At the stream level by providing `rate` attribute (integer value above zero).
- At the virtual sensor output level by providing `output-specification` → `rate` attribute (integer value above zero).

The rate control is a positive integer, and defines the minimum allowed time difference between two successive stream elements. For instance, if one is interested in receiving an average of a given sensor once an hour but the sensor underneath can produce arbitrary number of stream elements (e.g., due to uncontrollable packet losses in the internal network), he can express this behavior by setting the rate attribute of the virtual sensor output (`output-specification` → `rate`) to "3600000" (one hour is 3,600,000 milliseconds).

Refer to the virtual sensor quick reference for the syntactical information about different portions of the virtual sensor file.

2.5 GSN to GSN communication Protocol

In this section we would like to present the the low level details of GSN to GSN communication protocol. In order to enable data sharing and distributed collaborative data stream processing, we have introduced two special type of wrappers in GSN. First, the `local` wrapper, which enables data stream sharing among virtual sensors on the same machine. Second, the `remote` wrapper, which enables data stream sharing among multiple distributed virtual sensors each of which located on different machine accessible through the network.

2.5.1 `remote` wrapper

In GSN whenever a virtual sensor wants to use another virtual sensor located on a different GSN server, the communication between two GSN servers is triggered (during the loading process of the local virtual sensor). Once GSN notices that a remote virtual sensor is required by a local virtual sensor, GSN temporary suspends the local virtual sensor's loading process to confirm the existence of the remote virtual sensor. Therefore, GSN to GSN communication is initiated whenever a virtual sensor in a node A wants to use the data stream provided by another virtual sensor in a node B ($A \neq B$).

Using this kind of architecture, GSN mediates all the outgoing and incoming connections therefore the local virtual sensor does not interact directly with the remote virtual sensor (and vise versa). The packets exchanged between two GSN servers during GSN to GSN communication is depicted in figure 2.4 (all communications are implemented using XML-RPC calls). In the following we provide a brief description of each packet:

structure-request/structure is used by the local GSN server to discover the output structure of the remote virtual sensor. The response to this packet, confirms the existence and

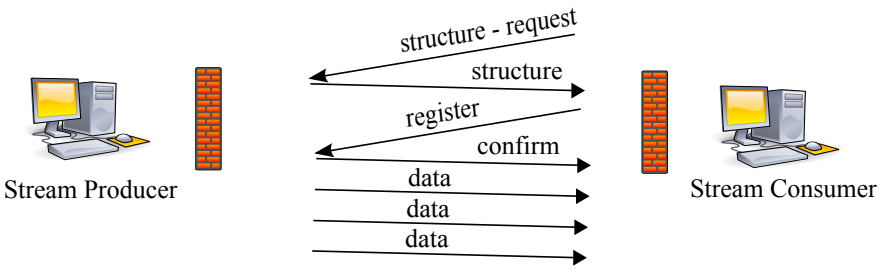

Figure 2.4: Experimental setup

availability of the remote virtual sensor and contains the details of the output-structure of the remote virtual sensor.

register/confirm is used by the local GSN server to send the query and the contact address of the stream consumer. The query will be added to the notification list associated with the prospective virtual sensor at the stream producer, therefore whenever the remote virtual sensor produces a stream element, the query will be evaluated and the output of the evaluation(in case it is not empty) is delivered to the stream consumer. The remote virtual sensor uses the addressing information (received in the registration packet) to contact the stream consumer in order to deliver the stream elements. As there might be multiple virtual sensors at the stream consumer side be interested in one virtual sensor hosted at the stream producer, any register request has a UUID associated with it which is used by the stream producer whenever it wants to deliver stream elements to the stream consumer.

data represents the stream of tuples which are going to be delivered to the stream consumer. At the stream consumer, GSN server receives the data and based on the UUID of the tuples, GSN server disseminates the tuples to the appropriate local virtual sensors.

In order to make the GSN to GSN communication more concrete, we provide more system level details below. For using a remote virtual sensor, the first step is locating the *contact point* of the GSN server which hosts the prospective virtual sensor. By default, the contact point is `http://ip-address:gsn-port/gsn-handler`[9][10]. If the contact point is correctly identified, the response to a plain HTTP POST request returns a XML output.[11][12]

Correct identification of the contact point is crucial in success of using the remote virtual sensor. Once the contact points identified successfully, one can define a stream which consume data from the other data source. Note that consuming data from a remote virtual sensor doesn't require any kind of modifications at the remote host and in fact due to GSN's decoupled architecture, the remote virtual sensor is not even aware of its data consumers. In figure 2.5, the virtual sensor `ConsumerVS` running at the GSN server with the IP address of `3.3.3.3` under the port `22001` is interested in getting data from the `CoolVS` running at the GSN server with the IP address of `4.4.4.4` under the port `22001`. To enable this communication one has to use a source configuration similar to the one in figure 2.5.

```
<address wrapper="remote">
```

[9]The GSN port is specified in the `conf/gsn.xml` file and will be 22001 unless changed.

[10]In the `webapp/WEB-INF/web.xml` file, the GSN's RPC handler (the `gsn.GSNRPC` class) is mapped to `/gsn-handler`. One shouldn't confuse the `/gsn-handler` with `/gsn` which is designed to be used solely by the web/ajax interface and does not involved in XML-RPC calls.

[11]The actual output represents an error as the request is not properly formatted.

[12]For sending plain HTTP POST requests to `http://ip-address:gsn-port/gsn-handler`, you may want to use http://code.google.com/p/rest-client/.

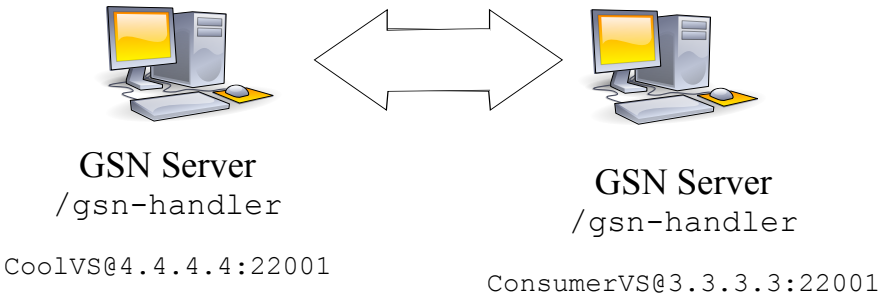

Figure 2.5: Simple GSN to GSN communication

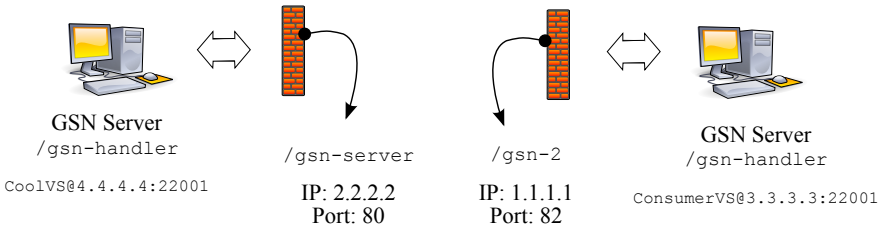

Figure 2.6: GSN to GSN communication with NAT

```
  <predicate key="name">CoolVS</predicate>
  <predicate key="host">4.4.4.4</predicate>
  <predicate key="port">22001</predicate>
</address>
```

Listing 2.5: Source configuration for simple GSN to GSN communication.

In some deployments, GSN servers are hosted behind a NAT an apache web server[13] which can cause port and/or IP change. This can be true for both the GSN data stream consumer and GSN data stream producer. In these cases, one can use the more advanced form of the remote wrapper. Figure 2.6 presents a sample setup in which both of the GSN data stream consumer and data producer are behind firewall. The firewall at the consumer side has mapped `3.3.3.3:22001` into `1.1.1.1:82` and at the stream producer side firewall has mapped `4.4.4.4:22001` into `2.2.2.2:80`. To enable this kind of communication one has to use a source configuration similar to the one in figure 2.6.

```
<address wrapper="remote">
  <predicate key="name">CoolVS</predicate>
  <predicate key="local-contact-point">http://1.1.1.1:82/gsn-2</predicate>
  <predicate key="remote-contact-point">http://2.2.2.2:80/gsn-server</predicate>
</address>
```

Listing 2.6: Source configuration for NATed GSN to GSN communication.

2.5.2 `local` wrapper

The `local` wrapper is special version of `remote` wrapper (host = "`127.0.0.1`") which is optimized for communication among two different virtual sensors inside the same GSN server. By having the `local` wrapper optimized, we imply that most of the overhead associated with TCP/IP networking calls are eliminated by using internal GSN calls instead. The `local`

[13]The instruction for using GSN behind a apache web server is provided in appendix refapp:gsn-apache.

wrapper is recommend whenever the end to end delay between two virtual sensors is important. In GSN, we have implemented the notification system so that the GSN server always gives priority to the local virtual sensors when it wans to disseminate the stream elements thus the local virtual sensors usually get notified earlier.

2.6 GSN Notifications

2.6.0.1 Introduction

In GSN, virtual sensors can be configured to notify users of certain events, e.g. to send an Email notification to an user informing them that a particular event has occurred. To implement notifications in GSN is very straight forward. The basic principle is that once the virtual sensor query is answered as specified in the virtual sensor description file, e.g.

```
<query>SELECT temperature FROM s1 WHERE temperature >= 100</query>
```

a notification can be triggered by the java processing class

```
<class-name>gsn.vsensor.EmailVirtualSensor</class-name>
```

see examples in next section. Thus, any type of notifications, e.g. Email, SMS, SIP, Fax, MMS can be implemented easily in a virtual sensor processing class.

The technical details of implementing notifications are left to the designer. Below are three examples of some of the notification services already implemented in GSN.

2.7 Implementation

The GSN implementation consists of the GSN-CORE, implemented in Java, and the platform-specific GSN-WRAPPERS, implemented in Java, C, and Ruby, depending on the available toolkits for accessing specific types of sensors or sensor networks. The implementation currently has approximately 80,000 lines of code and is available from SourceForge (http://gsn.sourceforge.net/). GSN is implemented to be highly modular in order to be deployable on various hardware platforms from workstations to small programmable PDAs, i.e., depending on the specific platforms only a subset of modules may be used. GSN also includes visualization systems for plotting data and visualizing the network structure. In the following sections we are going to discuss some of the key aspects of the GSN implementation

2.7.1 Adding new sensor platforms

For deploying a virtual sensor the user only has to specify an XML document as described in Section 2.2.1, if GSN already includes software support for the concerned hardware/software. Adding a new type of sensor or sensor network can be done by supplying the name of the wrapper (specified in `/conf/wrappers.properties`) conforming to the GSN API. At the moment GSN provides the following wrappers:

HTTP generic wrapper is used to pull data from devices via HTTP GET or POST requests, for example, the AXIS206W wireless camera.

TinyOS wrapper enables interaction with TinyOS compatible motes (version 1.x and 2.x). This wrapper uses the serial forwarder which is the standard access tool for TinyOS provided in the TinyOS package.

USB camera wrapper is used for dealing with cameras connected via USB to the local machine. As USB cameras are very cheap, they are quite popular as sensing devices. The wrapper supports cameras with OV518 and OV511 chips (see http://alpha.dyndns.org/ov511/).

TI-RFID wrapper enables access to Texas Instruments Series 6000 S6700 multi-protocol RFID readers.

Generic UDP wrapper can be used for any device using the UDP protocol to send data.

Generic serial wrapper supports sensing devices which send data through the serial port.

Additionally, we provide template implementations for standard cases and frequently used platforms. If wrapper implementations are shared publicly this also facilitates building a reusable code base for virtually any sensor platform. The effort to implement wrappers is quite low.

New wrappers can be added to GSN without having to rebuild or modify the GSN server (plug-and-play). Upon startup GSN locates the wrapper mappings through reading the `/conf/wrapper.properties` file and loads each wrapper whenever needed by the system.

2.7.2 Dynamic resource management

The highly dynamic processing environment we target with GSN requires adaptive dynamic resource management to allow the system to quickly react to changing processing needs and environmental conditions. Dynamic resource management accomplishes three main tasks:

Resource sharing: As the user can modify/remove/add virtual sensors on-the-fly during runtime, the system needs to keep track of all resources used by the individual virtual sensors and enforce resource sharing among sensors (wrappers) where possible.

Failure management: If GSN detects a faulty virtual sensor or wrapper, e.g., by runtime exceptions, GSN undeploys it and releases the associated resources.

Explicit resource control: The user can specify explicit memory and processing requirements and restrictions. While restrictions are always enforced, requirements are handled depending of the globally available resources of the GSN instance. GSN tries to share the available resources in a fair way taking into account the explicitly specified resource requirements, if provided.

Dynamic resource management is performed at several levels in GSN as shown in Figure 2.7. Separating the resource sharing into several layers logically decouples the requirements and allows us to achieve a higher level of reuse of resources. In the following we will discuss the different levels.

Wrapper sharing. Wrappers communicate directly with the sensors which involves expensive I/O operations via a serial connection or wireless/wired network communication. To minimize the costs incurred by these operations GSN shares wrappers among virtual sensors accessing the same physical/virtual sensors. To do so each GSN node maintains a repository of active wrappers. If a new virtual sensor is deployed, the node first checks with the wrapper

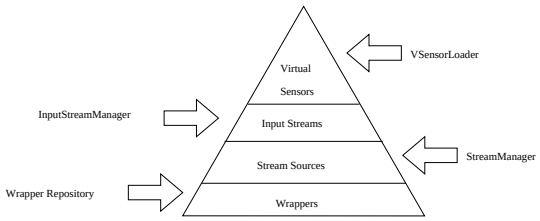

Figure 2.7: Hierarchical resource sharing in GSN

repository whether an identical wrapper already exists, i.e., wrapper name and initialization parameters (and their corresponding values) of the `<wrapper>` element in the virtual sensor definitions are identical. If a match is found, the new virtual sensor is registered to the existing wrapper as a consumer. If not, a new wrapper instance is created and registered with the wrapper repository. In the case of remote sensor accesses this strategy is applied at both the sending and receiving sides to maximize the sharing, i.e., multiple virtual sensors on one GSN node share a wrapper for the same remote sensor and on the node hosting the sensor the wrapper is shared among all nodes accessing it.

Data sharing. The raw input data produced by the wrappers is processed and filtered by the source queries to generate the actual input data for the input streams of a virtual sensor. For this purpose a source defines what part of the raw input data is used by the associated source query to produce the source's output data, i.e., by defining the available storage, sampling rates, and window sizes a view on the raw data is defined on which the source query is executed. In terms of the implementation each wrapper is assigned a storage holding the raw data and source queries are then defined as *SQL views* on this data store.

This has a number of advantages: (1) It minimizes the storage consumption as raw data is only stored once. Especially if the sensor data is large, e.g., image data, this is relevant. (2) If the sensor data comes from a power-constrained or slow device, power is conserved and processing is sped up. (3) Different processing strategies can be applied to the same data without having to replicate it, for example, image enhancement algorithms and object detection can use the same raw image data.

In the same way as a wrapper can be shared by multiple sources, a source can also be shared among multiple streams at a higher level, and streams in turn are shared by multiple virtual sensors. In essence each of the layers in Figure 2.7 can be viewed as a resource pool where each of the individual resources in the pool can be shared among multiple resources at the next higher level. Conversely, each higher level resource can also use any number of lower level resources.

2.7.3 Query planning and execution

In GSN each virtual sensor corresponds to a database table and each sensor reading corresponds to a new tuple in the related table. As we use a standard SQL database as our low-level query processing engine, the question is how to represent the streaming logic in a form understandable for a standard database engine (as already described, GSN separates the stream processing directives from the query). We address this problem by using a query translator which gets an SQL query and the stream processing directives as provided in the virtual sensor definition as inputs and translates this into a query executable in a standard database. The query translator relies on special support functions which emulate stream-oriented constructs in a database. These support functions are dependent on the database used and are provided by

GSN (currently we provide adapters for H2 and MySQL). Translated queries are cached for subsequent use.

Upon deployment of a virtual sensor VS, all queries Q_i contained in its specification are extracted. Each query $Q_i(VS_1, \ldots, VS_n)$ accesses one or more relations $VS_1, \ldots, VS_n$ which correspond to virtual sensors. Then the query translator translates each $Q_i(VS_1, \ldots, VS_n)$ into an executable query $Q_i^t(VS_1, \ldots, VS_n)$ as described above and each $Q_i^t(VS_1, \ldots, VS_n)$ is declared as a view in the database with a unique identifier Id_i. This means whenever a new tuple, i.e., sensor reading, is added to the database, the concerned views will automatically be updated by the database. Additionally, a tuple (VS_j, Id_i, VS) for each $VS_j \in VS_1, \ldots, VS_n$ is added to a special view registration table. This procedure is done once when a virtual sensor is deployed.

With this setup it is now simple to execute queries over the data streams produced by virtual sensors: As soon a new sensor reading for a virtual sensor VS_d becomes available, it is entered into the appropriate database relation. Then the database server queries the registration table using VS_d as the key and gets all identifiers Id_r registered for new data of VS_d. Then simply all views V_r affected by the new data item can be retrieved using the Id_r and all V_r can be queried using a `SELECT * FROM V_r` statement and the resulting data can be returned to the virtual sensor containing V_r (third column in the registration table). Since views are automatically updated by the database querying them is efficient. However, with many registered views (thousands or more) scalability may suffer. Thus GSN does not produce an individual query for each view but merges all queries into a large select statement, and the result will then be joined with the view registration table on the view identifier. Thus the result will hold tuples that identify the virtual sensor to notify of the new data. The reasons for applying this strategy are that (1) database connections are expensive, (2) with increasing number of clients and virtual sensor definitions, the probability of overlaps in the result sets increases which automatically will be exploited by the database's query processor, and (3) query execution in the database is expensive, so one large query is much less costly than many (possibly thousands) small ones.

Immediate notification of new sensor data is currently implemented in GSN and is an eager strategy. As an alternative also a lazy strategy could be used where the query execution would only take place when the GSN instance requests it from the database, for example, periodically at regular intervals. In practice the former can be implemented using views or triggers and the latter can be implemented using inner selects or stored procedures.

2.7.4 Network communication

Looking inside the GSN infrastructure, there are at least half a dozen difference network communication channels are used. In this section I would like to dive in to the details of the some major communication protocols designed and implemented in GSN.

2.7.4.1 Reusing Data Streams

One of the main ideas behind the virtual sensors is resuability. The resuability comes in two forms. First being able to recreate the same processing logic on different data streams. Second being able to reuse streaming data produced by other parties over the internet and possibly create a new data stream but instrumenting the original streams. In this section, I present the both high level and low level details associated with the second aspect of the reusability.

The virtual sensor descriptor file is the first place which specifies the intention of reusing streaming data from another virtual sensor. The source virtual sensor can be located anywhere

as long as it is accessible through the network, this ofcourse includes the local machine and any other machine on the Internet.

In GSN, our vision is having an internet scale streaming world in which people can publish streaming data which can be produced directly using some sort of a measurement device which can range from a physical wireless sensor to stock ticks from a financial market.

List of Figures

List of Tables

Listings

Bibliography

[1] Daniel J. Abadi, Yanif Ahmad, Magdalena Balazinska, Ugur Çetintemel, Mitch Cherniack, Jeong-Hyon Hwang, Wolfgang Lindner, Anurag Maskey, Alex Rasin, Esther Ryvkina, Nesime Tatbul, Ying Xing, and Stanley B. Zdonik. The Design of the Borealis Stream Processing Engine. In *CIDR*, 2005. 43

[2] A. Arasu, B. Babcock, S. Babu, J. Cieslewicz, M. Datar, K. Ito, R. Motwani, U. Srivastava, and J. Widom. *Data-Stream Management: Processing High-Speed Data Streams*, chapter STREAM: The Stanford Data Stream Management System. Springer, 2006. 43

[3] Sirish Chandrasekaran, Owen Cooper, Amol Deshpande, Michael J. Franklin, Joseph M. Hellerstein, Wei Hong, Sailesh Krishnamurthy, Samuel Madden, Vijayshankar Raman, Frederick Reiss, and Mehul A. Shah. TelegraphCQ: Continuous Dataflow Processing for an Uncertain World. In *CIDR*, 2003. 43

[4] Mitch Cherniack, Hari Balakrishnan, Magdalena Balazinska, Donald Carney, Ugur Çetintemel, Ying Xing, and Stanley B. Zdonik. Scalable Distributed Stream Processing. In *CIDR*, 2003. 43

[5] M. Franklin, S. Jeffery, S. Krishnamurthy, F. Reiss, S. Rizvi, E. Wu, O. Cooper, A. Edakkunni, and W. Hong. Design Considerations for High Fan-in Systems: The HiFi Approach. In *CIDR*, 2005. 42

[6] P. B. Gibbons, B. Karp, Y. Ke, S. Nath, and S. Seshan. IrisNet: An Architecture for a World-Wide Sensor Web. *IEEE Pervasive Computing*, 2(4), 2003. 42

[7] A. J. G. Gray and W. Nutt. A Data Stream Publish/Subscribe Architecture with Self-adapting Queries. In *International Conference on Cooperative Information Systems (CoopIS)*, 2005. 43

[8] Sean Rooney, Daniel Bauer, and Paolo Scotton. Techniques for Integrating Sensors into the Enterprise Network. *IEEE eTransactions on Network and Service Management*, 2(1), 2006. 43

[9] firstname1 secondname1 and firstname2 secondname2. Dummy Article. *Dummy Journal*, 20 August 2008.

[10] M. Sgroi, A. Wolisz, A. Sangiovanni-Vincentelli, and J. M. Rabaey. A service-based universal application interface for ad hoc wireless sensor and actuator networks. In *Ambient Intelligence*. Springer Verlag, 2005. 42

[11] J. Shneidman, P. Pietzuch, J. Ledlie, M. Roussopoulos, M. Seltzer, and M. Welsh. Hourglass: An Infrastructure for Connecting Sensor Networks and Applications. Technical Report TR-21-04, Harvard University, EECS, 2004. http://www.eecs.harvard.edu/~syrah/hourglass/papers/tr2104.pdf. 42

[12] Yong Yao and Johannes Gehrke. Query Processing in Sensor Networks. In *CIDR*, 2003. 43

[13] Stan Zdonik, Michael Stonebraker, Mitch Cherniack, Ugur Cetintemel, Magdalena Balazinska, and Hari Balakrishnan. The Aurora and Medusa Projects. *Bulletin of the Technical Committe on Data Engineering, IEEE Computer Society*, 2003. 43

Appendix A

Quick Reference Guide

A.1 Virtual Sensors (*VS*)

A.1.1 *VSD* DTD

All the *VS* are configured with an XML Virtual Sensor Description file (*VSD*). A graphical representation of the *VSD* Document Type Definition (DTD) is available on the Figure A.1. The description of all these tags are given in the Table A.1.

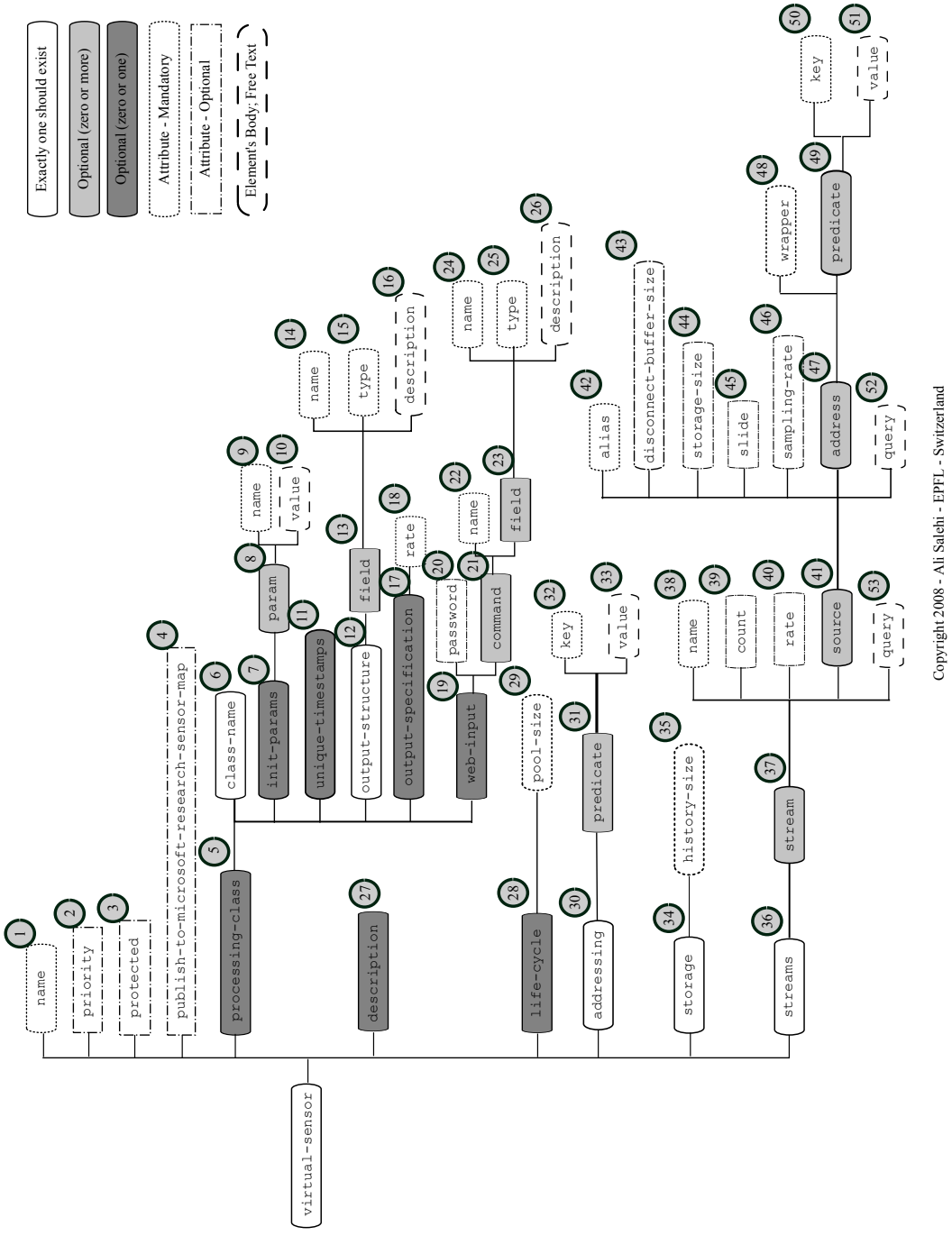

Figure A.1: *VSD* DTD Quick Reference Card

VSD DTD Quick Reference Card Description				
Tag	**Name**	**Mandatory**	**Allowed Values**	**Description**
1	name	1 only	alpha	Identifies the VS and must be unique in an instance of GSN
2	protected	0 or 1	TODO	TODO
3	priority	0 or 1	0 - 20	Defaults to 10, 0 is highest and 20 lowest
4	publish-to-microsoft-research-map	0 or 1		
5	processing-class			Container for *VSP* specification
6	class-name	1 only	valid path	Path to Java implementation of *VSP*
7 - 10	init-params	0 or 1		Parameter specific to *VSP* (Refer to section **??**)
11	unique-timestamps	0 or 1	true/false	Unless set to false duplicate timestamps are dropped with a warning
12	output-structure	1 only		Sensor data structure
13	field	0 or more		field definition (Name and type must match source query
14	name	1 per field	alpha	unique field name
15	type	1 per field	valid type	Field data type
16	description	1 per field	Free text	Description of output item
17 , 18	output-specification rate	1 only	numeric	Minimum interval between generated data items.
19	web-input	0 or 1	TODO	TODO
20	password	0 or 1	TODO	TODO
21	command	0 or more	TODO	TODO
22	name	1 only	TODO	TODO
23	field	0 or more	TODO	TODO
24	name	1 only	TODO	TODO
25	type	1 only	TODO	TODO
26	description	1 only	Free text	Field description
27	description	0 or 1	Free text	*VS* Description
28, 29	life-cycle pool-size	1 only		A performance parameter set maximum number of instances
30 - 33	addressing		key/value pairs	Predicates describing sensor location
34,35	storage history-size	1 only	output-specification	Number of stream elements held in database - does not affect stream processing
36	streams	1 only		Container for streams
37	stream	0 or more		Container for a single stream definition
38	name	1 only	alhanumeric	Unique stream identifier
39	count	0 or 1		deprecated ?
40	rate	1 only	numeric (ms)	performance tuning - minimum interval (ms.) between calls to this stream
41	source	0 or more		Container for one source
42	alias	1 only	alphanumeric	Unique source identifier
43	disconnect-buffer-size	0 or 1	TODO	TODO
44	storage-size	0 or 1		Size of sliding window (count or time based)
45	slide	0 or 1		Slide interval (count or time based)
46	sampling-rate	0 or 1	float 0 - 1	load shedding parameter
47	address	0 or 1		Container to contain wrappers
48	wrapper	1 or more	alphanumeric	Wrapper name
49 - 51	predicate	0 or more		Predicates are specific to wrappers (Refer to section **??**)
52	query (source)	1 only	Valid SQL	Selects data from source
53	query (stream)	1 only	Valid SQL	Selects data from stream

Table A.1: *VSD* DTD Quick Reference Card Description

A.2 GSN ANT Tasks

GSN ANT Tasks	
Task Name	**Description**
start-all	Start both the Safe Storage and the GSN processes.
stop-all	Stop both the Safe Storage and the GSN processes.
start-acquisition	Start the Safe Storage process. The wrapper that were loaded during the last runs will be automatically resumed and will directly start acquiring data.
clean-acquisition	Delete all the Safe Storage permanent storage and flush the list of Wrappers to resume. Use this task with caution since it may delete some unprocessed data forever.
stop-acquisition	Stop the Safe Storage process.
gsn	Start the GSN process. You also have to start the Safe Storage process if you are using Safe Storage wrappers.
stop	Stop the GSN process. The Safe Storage processes if any will continue running and acquire the data.
restart	Stop and restart the GSN process.
gui	Starts the GSN graphical user interface.
jar	Creates a jar _le from the source.
clean	Removes the current build files and forces a rebuild.
cleandb	Removes the redundant tables which are create for holding GSN's internal states.
compile-reports	Compile the Jasper Reports located in the `gsn-reports` directory. Must be called after modification of any Jasper report configuration file (.jrxml).
Use each of these tasks by typing in your terminal: `ant <Task Name>`	

Table A.2: GSN ANT Tasks

Appendix B

GSN Tutorials

B.1 Understanding GSN Virtual Sensors

The internal behaviour of GSN is of a content-based publish-subscribe system, on which subscribers (virtual sensors) are subscribed (using SQL queries) to publishers (wrappers). Content (sensor data) is described as timestamped tuples and stored in tables in a relational database. GSN handles the storage and event notifications internally.

All this is implemented in GSN as follows:

1. When a wrapper is loaded, a table is created and given a random name. The table's fields are obtained from the DataField[] array defined in the wrapper code. This table will be use to store any data coming from the sensor and it will be automatically timestamped.
2. When a virtual-sensor is loaded, several things happen. First, a view[1] is created from the wrapper table based on the SQL query specified in the <source> section in the virtual-sensor XML file[2]. The SQL query in this section is used to "select" which data fields from the wrapper table are selected for the view. This step is repeated for any number of data sources declared in the virtual-sensor XML file. Second, once the view is created (internally), a table is created for the virtual-sensor, the size of this table is declared in the virtual-sensor XML file in the <storage> section, and the name of the table is the name given to the virtual-sensor. The data fields of this table are obtained from the <output-structure> section in the virtual-sensor XML file and they have to match with the SQL query section <stream> section[3].

Note that most of this is hidden from the user, in this document we will describe how this is implemented in GSN.

A **Virtual Sensor** (*VS*) is the main component in gsn. It receives data from one or more $Wrapper(s)$. It can combine their data, process and finally store it. A *VS* is defined in a single Virtual Sensor Description file (*VSD)* and combines different pieces of software

[1] http://en.wikipedia.org/wiki/Database_view

[2] In GSN terms, this is know as the "source query".

[3] This is know as the "virtual sensor query".

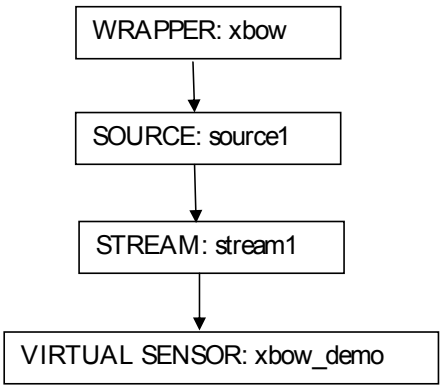

Figure B.1:

- One Virtual Sensor Processing Class (VSP)
- Zero or Many $Wrapper(s)$

Depending on the actual virtual sensor used, the processed data may be published as a data stream, displayed as a chart, sent to a database or used in any number of ways limited only by the ingenuity of the developer.

The GSN software includes a library of standard virtual sensors and a framework that facilitates the development of others. The configuration of any specific instance of a virtual sensor is defined in an XML document called a virtual sensor definition (VSD).

We use the multiFormatSample virtual sensor, a fairly basic instance of a virtual sensor to demonstrate and follow the flow of data using the Bridge Virtual Sensor. The flow of data in this basic sensor is depicted in the following illustration:

B.1.1 The multiFormatSample Virtual Sensor.

We will use the MultiFormat wrapper (src.gsn.wrappers.MultiFormatWrapper.java) since it is the simplest wrapper to study (and modify) and the wrapper "simulates" a sensor which provides temperature, light, and packet type sensor readings every second but does not actually rely on an external data source for its operation.

B.1.2 Virtual Sensor Description File

A **Virtual Sensor Description file** (VSD) is an XML file that contains the selection and the parameterization of the VSP and wrapper that compose a VS. This file also contains the SQL statements that connect them together.

The multiFormatSample Virtual Sensor is defined by the following virtual-sensor XML file:

```
<virtual-sensor name="MultiFormatTemperatureHandler" priority="10">
   <processing-class>
      <class-name>gsn.vsensor.LightVirtualSensor</class-name>
      <init-params />
      <output-structure>
         <field name="light" type="double"/>
         <field name="temperature" type="double"/>
         <field name="packet_type" type="int/>
```

```
      </output-structure>
   </processing-class>
   <description>Simulates sensor readings every second.</description>
   <life-cycle pool-size="10" />
   <addressing>
      <predicate key="geographical">Sensor 114 @ EPFL</predicate>
      <predicate key="LATITUDE">46.520000</predicate>
      <predicate key="LONGITUDE">6.565000</predicate>
   </addressing>
   <storage history-size="10"/>
   <streams>
      <stream name="input1">
         <source alias="source1" sampling-rate="1" storage-size="1">
            <address wrapper="multiformat">
               <predicate key="HOST">localhost</predicate>
               <predicate key="PORT">22001</predicate>
            </address>
            <query>SELECT light, temperature, timed FROM wrapper</query>
         </source>
         <query>source1.light AS light_sensorFROM source1</query>
      </stream>
   </streams>
</virtual-sensor>
```

Listing B.1: multiFormatSample.xml

Listing multiFormatSample.xml

B.1.3 Wrapper

A GSN Wrapper (Wrapper) is a piece of Java code that does the data acquisition for a specific type of device..

The wrapper from which the data will be selected is defined in the address element of the source, for example in multiFormatWrapper:

```
<address wrapper="multiformat">
   <predicate key="HOST">localhost</predicate>
   <predicate key="PORT">22001</predicate>
</address>
```

The predicates included in the address element are parameters that are particular to the wrapper used. As the multiFormat wrapper generates its own data the address predicates are not really required but are included for illustration only.

In a wrapper that connects to an external data source, the predicates contain parameters necessary to identify the instance of the wrapper to be used. In this case, if the wrapper connected to a real data sourse, the parameters would define a network host address and port to connect to the sensor.

The wrapper creates a table from the DataField[] structure:

```
private DataField[] collection = new DataField[] {
   new DataField("packet_type", "int", "packet type"),
   new DataField("temperature", "double", "Presents the temperature sensor."),
   new DataField("light", "double", "Presents the light sensor.") };
```

Listing B.2: DataField declaration in multiFormatWrapper.java

When GSN is started, wrappers tables are given randomly generated names, for example, the table for the MultiFormatWrapper in this instance is given the name _501577155.

If we query that table (see), we notice the following fields: timed, PACKET_TYPE, TEMPERATURE, LIGHT, and a primary key field PK. Some of the data fields are automatically converted to upper case by GSN for clarity. These fields correspond to the ones defined in the DataField[]array in the wrapper code:

The timed field is automatically generated by GSN to indicate the timestamp of the tuple. The timestamp is expressed in unix epoch time.

multiFormat Wrapper data table				
PK	**timed**	**PACKET_TYPE**	**TEMPERATURE**	**LIGHT**
386	1225841381231	1	NULL	779
387	1225841382591	1	NULL	329.2

Table B.1: multiFormat Wrapper data table

Once the wrapper code is initialized and the wrapper table created then when a virtual-sensor is loaded, a view or a set of views is created to represent the data source

B.1.4 Source

The source element of the VSD includes the wrapper declaration and an SQL query to select the required data:

```
<source alias="source1" sampling-rate="1" storage-size="1">
      <address wrapper="multiformat">
         <predicate key="HOST">localhost</predicate>
         <predicate key="PORT">22001</predicate>
      </address>
      <query>SELECT light, temperature, timed FROM wrapper</query>
   </source>
```

The attributes of the source include:

alias an identifier for this source

storage-size the window size - the number of data items stored in the temporary table which sets the number of record used to calculate aggregate functions such as AVG, MIN, MAX or SUM. In the example, no aggregate is used so the value is set at 1. (In the sample, two rows are present because the sample was viewed after a new record was recieved but before the previous one was dropped.

slide the amount by which the sliding window moves when generating aggregate queries. For example, a slide value of 10 would generate a query on the arrival of each tenth record. (slide is not relevant to queries not using aggregate funcions such as the present example.)

sampling-rate provides for load shedding when the source generates data at a higher rate than required. Has a value between 0 and 1. A value of 1 means no data is dropped, while a value of 0.2 means only 20% of data items will be processed and the remainder (80%) will be silently dropped.

The data view for source1 is named _695753603and contains the following data fields: light, temperature, packet type and timed (see).

Since views are virtual tables created by stored queries. This view corresponds to the "source" query defined in the virtual-sensor XML file (). In this way, if several data sources are declared, GSN can use views to "merge" data from wrapper tables, in other words, views act as temporary tables to hold data used by virtual-sensors.

SELECT * FROM _695753603;			
light	**temperature**	**packet_type**	**timed**
329.2	NULL	1	1225841382591

Table B.2: Source data view

B.1.5 Stream

A stream tells GSN what data to send to the result table created for the virtual sensor. The stream declaration defines the source views from which data will be selected and the SQL query to select data to be added to the table:

```
<stream name="input1">
   <source alias="source1" sampling-rate="1" storage-size="1">
      <address wrapper="multiformat">
         <predicate key="HOST">localhost</predicate>
         <predicate key="PORT">22001</predicate>
      </address>
      <query>SELECT light, temperature, timed FROM wrapper</query>
   </source>
   <query>source1.light AS light_sensorFROM source1</query>
</stream>
```

The attributes of a stream include:

name a mandatory identifier for the stream.

rate The rate parameter is a performance tuning parameter. It defines the minimum interval in milliseconds between two calls to this virtual sensor. If there is data available for the virtual sensor in less than this value, then the data is silently dropped (Book of GSN p. 25).

count Count puts a limit on the life cycle of the stream query in terms of number of outputs.

For instance, if the count is 100 it implies that the stream source should become disabled after producing 100 values (stream elements). This property is used rarely and it is planned to be removed from the next release.

The stream query selects data from one or more sources and can perform joins to combine data from multiple sources. The result set of the query is inserted into a table bearing the name of the virtual sensor which becomes the output of the sensor. The 'timed' field can be selected from one of the sources or failing that the time at the GSN container host will be used.

SELECT * FROM multiformattemperaturehandler;				
PK	**timed**	**LIGHT**	**TEMPERATURE**	**PACKET_TYPE**
74552	1225841363432	637.9	NULL	1
74553	1225841364792	507.3	60.2	2
74554	1225841368402	NULL	17.3	2
74555	1225841369042	691.9	NULL	1
74556	1225841372136	380.3	NULL	1
74557	1225841375230	834.5	NULL	1
74558	1225841376059	NULL	60.1	2
74559	1225841378168	NULL	75.7	2
74560	1225841381231	779	NULL	1
74561	1225841382591	329.2	NULL	1

Table B.3: multiFormat VS Output Table

B.1.6 Virtual Sensor

The root element of the VSD is labelled <virtual-sensor> and contains attributes that define the whole VS

name is a mandatory and arbitrary identifier for the VS which can be used by another virtual sensor to idntify this virtual sensor as a source using the remote or local virtual sensor class.

priority is optional and must be between 0 which is the highest priority and 20 is the lowest. The default priority is 10.

B.1.6.1 Virtual Sensor Processing Class

A **Virtual Sensor Processing class** (*vsp*) is a piece of Java code that process and stores the data upon reception from the wrapper.

```
<processing-class>
   <class-name>gsn.vsensor.LightVirtualSensor</class-name>
   <init-params />
   <output-structure>
      <field name="light" type="double"/>
      <field name="temperature" type="double"/>
      <field name="packet_type" type="int/>
   </output-structure>
</processing-class>
```

The key elements in this section are:

class-name specifies the name of a java class that implements the virtual sensor processing class (VSP). (gsn.vsensor.BridgeVirtualSensor is used in the example). The GSN distribution includes several VSP's and a framework in which others can can be developed.

unique-timestamps allows the virtual sensor to override the default which is to make the index to the timestamp UNIQUE. This would allow duplicated timestamps to be added to the table.

init-params allows for the provision of parameters to the virtual sensor class. Parameters are specific to individual sensor classes.

output-specification-rate acts in a similar manner to the rate attribute of a stream but presumably controls the the whole VS rather than an individual stream. In a VS with only one stream both seem to have the same effect which is to pevent the generation of another data row until the specified time (in milliseconds) has elapsed.

output-structure provides the structure of a row in the output data. The name and type attributes of each field element must match those of a data item selected in the stream query. The text contents of each field element can contain an optional description of the field.

B.1.6.2 Other Elements of a VSD

Other elements of the *VSD* include:

description can contain a textual description of the sensor.

life-cycle pool-size is a performance parameter. It is usually safe to keep the default value. It defines the maximum number of instances of this virtual sensor (with this configuration). This can happen when the processing method of the virtual sensor takes a long time to complete,

and / or when data arrives at high speed. If all instances are busy, then the data will be dropped.

The *addressing* element contains predicates which can be used to specify the location and other characteristics of the virtual sensor. The predefined keys "LATITUDE" and "LONGITUDE" can be used to locate the sensor on the map pages of the web interface.

storage history-size sets the number of records to be retained in the virtual sensor table. It does not impact the logical processing of data streams.

B.1.7 Summary

The multiFormatSample VS is a very simple example which simply selects data fields in the stream from a single wrapper. The examples were generated using release 890 of GSN and MySQL as the underlying database.

The real power of GSN lays in its ability to use more complex SQL queries with multiple data streams, sources and wrappers but that can only be approached with any confidence once the basics are understood.

Appendix C

GSN an Evaluation

C.1 Evaluation [1]

GSN aims at providing a zero-programming and efficient infrastructure for large-scale interconnected sensor networks. To justify this claim we experimentally evaluate the throughput of the local sensor data processing and the performance and scalability of query processing as the key influencing factors. As virtual sensors are addressed explicitly and GSN nodes communicate directly in a point-to-point (peer-to-peer) style, we can reasonably extrapolate the experimental results presented in this section to larger network sizes. For our experiments, we used the setup shown in Figure C.1.

The GSN network consisted of 5 standard Dell desktop PCs with Pentium 4, 3.2GHz Intel processors with 1MB cache, 1GB memory, 100Mbit Ethernet, running Debian 3.1 Linux with an unmodified kernel 2.4.27. For the storage layer use standard MySQL 5.18. The PCs were attached to the following sensor networks as shown in Figure C.1.

- A sensor network consisting of 10 Mica2 motes, each mote being equipped with light and temperature sensors. The packet size was configured to 15 Bytes (data portion excluding the headers).
- A sensor network consisting of 8 Mica2 motes, each equipped with light, temperature, acceleration, and sound sensors. The packet size was configured to 100 Bytes (data portion excluding the headers). The maximum possible packet size for TinyOS 1.x packets of the current TinyOS implementation is 128 bytes (including headers).
- A sensor network consisting of 4 Tiny-Nodes (TinyOS compatible motes produced by Shockfish, http://www.shockfish.com/), each equipped with a light and two temperature sensors with TinyOS standard packet size of 29 Bytes.
- 15 Wireless network cameras (AXIS 206W) which can capture 640x480 JPEG pictures with a rate of 30 frames per second. 5 cameras use the highest available compression (16kB average image size), 5 use medium compression (32kB average image size), and 5 use no compression (75kB average image size). The cameras are connected to a Linksys WRT54G wireless access point via 802.11b and the access point is connected via 100Mbit Ethernet to a GSN node.

[1] The evaluation results in this section correspond to GSN release 0.90

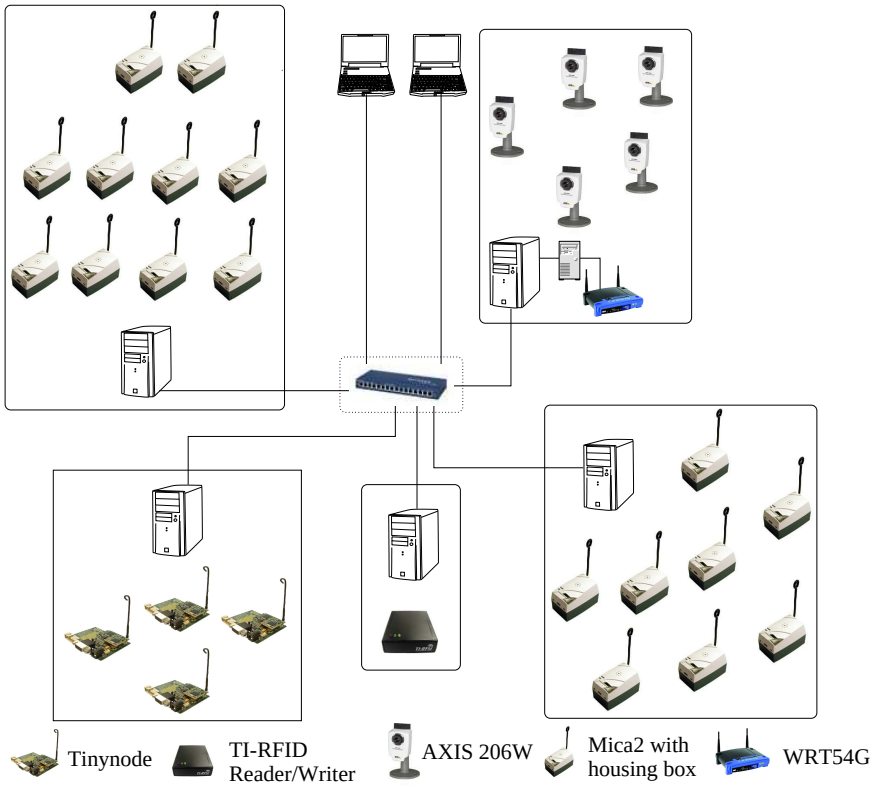

Figure C.1: Experimental setup

- A Texas Instruments Series 6000 S6700 multi-protocol RFID reader with three different kind of tags, which can keep up to 8KB of data. 128 Bytes capacity.

The motes in each sensor network form a sensor network and routing among the motes is done with the surge multi-hop ad-hoc routing algorithm provided by TinyOS.

C.1.1 Internal processing time

In the first experiment we wanted to determine the internal processing time a GSN node requires for processing sensor readings, i.e., the time interval when the wrapper gets the sensor data until the data can be provided to clients by the associated virtual sensor. This delay depends on the size of the sensor data and the rate at which the data is produced, but is independent of the number of clients wanting to receive the sensor data. Thus it is a lower bound and characterizes the efficiency of the implementation.

We configured the 22 motes and 15 cameras to produce data every 10, 25, 50, 100, 250, 500, and 1000 milliseconds. As the cameras have a maximum rate of 30 frames/second, i.e., a frame every 33 milliseconds, we added a proxy between the GSN node and the WRT54G access point which repeated the last available frame in order to reach a frame interval of 10 milliseconds. All GSN instances used the Sun Java Virtual Machine (1.5.0 update 6) with memory restricted to 64MB.

The experiment was conducted as follows: All motes and cameras were set to the same rate and produced data for 8 hours and we measured the processing delay. This was repeated 3

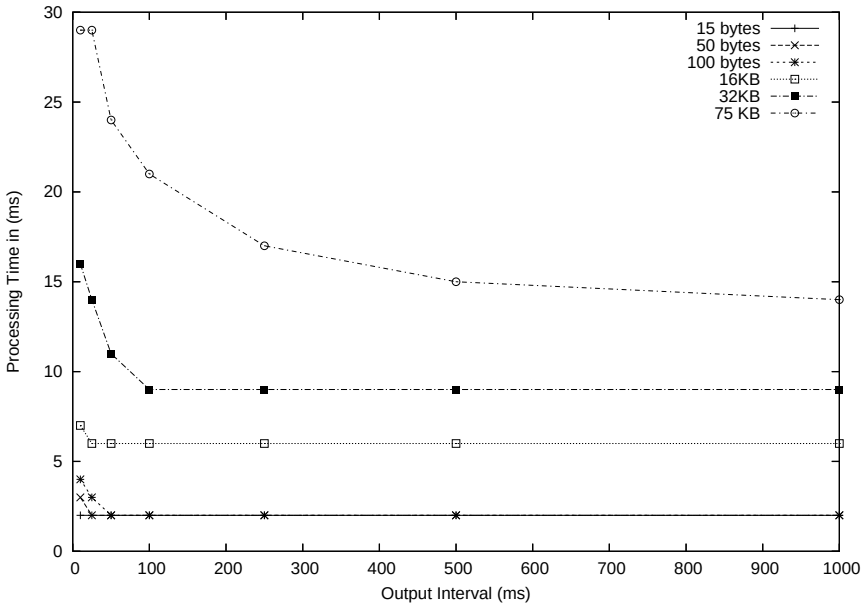

Figure C.2: GSN node under time-triggered load

times for each rate and the measurements were averaged. Figure C.2 shows the results of the experiment for the different data sizes produced by the motes and the cameras.

High data rates put some stress on the system but the absolute delays are still quite tolerable. The delays drop sharply if the interval is increased and then converge to a nearly constant time at a rate of approximately 4 readings/second or less. This result shows that GSN can tolerate high rates and incurs low overhead for realistic rates as in practical sensor deployments lower rates are more probable due to energy constraints of the sensor devices while still being able to deal also with high rates.

C.1.2 Scalability in the number of queries and clients

In this experiment the goal was to measure GSN's scalability in the number of clients and queries. To do so, we used two 1.8 GHz Centrino laptops with 1GB memory as shown in Figure C.1 which each ran 250 lightweight GSN instances. The lightweight GSN instance only included those components that we needed for the experiment. Each GSN-light instance used a random query generator to generate queries with varying table names, varying filtering condition complexity, and varying configuration parameters such as history size, sampling rate, etc. For the experiments we configured the query generator to produce random queries with 3 filtering predicates in the `where` clause on average, using random history sizes from 1 second up to 30 minutes and uniformly distributed random sampling rates (seconds) in the interval $[0.01, 1]$.

Then we configured the motes such that they produce a measurement each second but would deliver it with a probability $P < 1$, i.e., a reading would be dropped with probability $1 - P > 0$. Additionally, each mote could produce a burst of R readings at the highest possible speed depending on the hardware with probability $B > 0$, where R is a uniformly random integer from the interval $[1, 100]$. I.e., a burst would occur with a probability of $P * B$ and would produce randomly 1 up to 100 data items. In the experiments we used $P = 0.85$ and

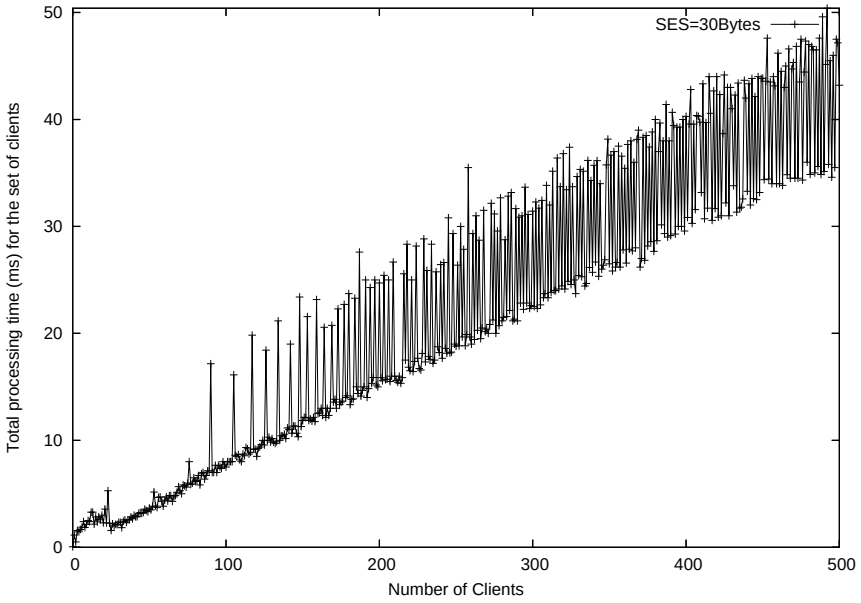

Figure C.3: Query processing latencies in a node

$B = 0.3$. On the desktops we used MySQL as the database with the recommended configuration for large memory systems. Figure C.3 shows the results for a stream element size (SES) of 30 Bytes. Using SES=32KB gives the same latencies. Due to space limitations we do not include this figure.

The spikes in the graphs are bursts as described above. Basically this experiment measures the performance of the database server under various loads which heavily depends on the used database. As expected the database server's performance is directly related to the number of the clients as with the increasing number of clients more queries are sent to the database and also the cost of the query compiling increases. Nevertheless, the query processing time is reasonably low as the graphs show that the average time to process a query if 500 clients issue queries is less than 50ms, i.e., approximately 0.5ms per client. If required, a cluster could be used to the improve query processing times which is supported by most of the existing databases already.

In the next experiment shown in Figure C.4 we look at the average processing time for a client excluding the query processing part. In this experiment we used $P = 0.85$, $B = 0.05$, and R is as above.

We can make three interesting observations from Figure C.4:

1. GSN only allocates resources for virtual sensors that are being used. The left side of the graph shows the situation when the first clients arrive and use virtual sensors. The system has to instantiate the virtual sensor and activates the necessary resources for query processing, notification, connection caching, etc. Thus for the first clients to arrive average processing times are a bit higher. CPU usage is around 34% in this interval. After a short time (around 30 clients) the initialization phase is over and the average processing time decreases as the newly arriving clients can already use the services in place. CPU usage then drops to around 12%.

2. Again the spikes in the graph relate to bursts. Although the processing time increases

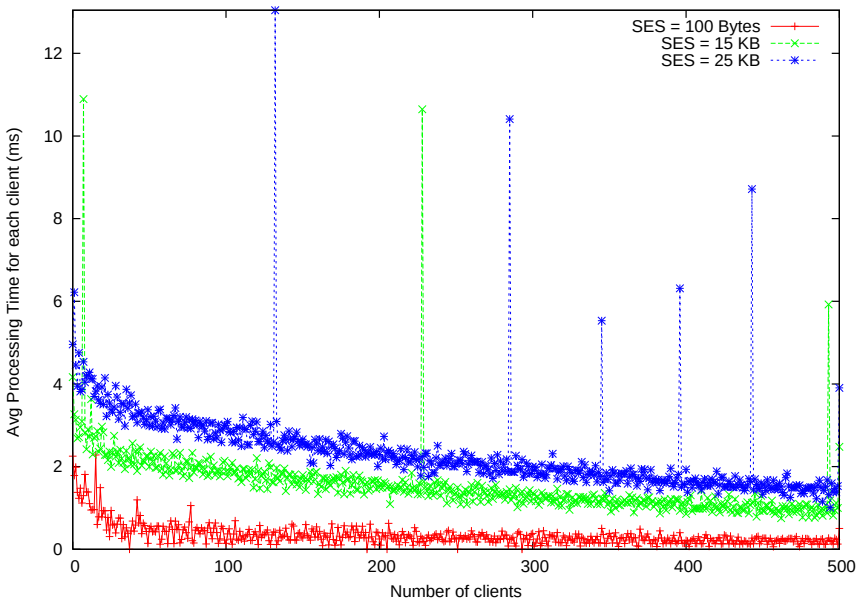

Figure C.4: Processing time per client

considerably during the bursts, the system immediately restores its normal behavior with low processing times when the bursts are over, i.e., it is very responsive and quickly adopts to varying loads.

3. As the number of clients increases, the average processing time for each client decreases. This is due to the implemented data sharing functionalities. As the number of clients increases, also the probability of using common resources and data items grows.

C.2 Related work

So far only few architectures to support interconnected sensor networks exist. Sgroi et al. [10] suggest basic abstractions, a standard set of services, and an API to free application developers from the details of the underlying sensor networks. However, the focus is on systematic definition and classification of abstractions and services, while GSN takes a more general view and provides not only APIs but a complete query processing and management infrastructure with a declarative language interface.

Hourglass [11] provides an Internet-based infrastructure for connecting sensor networks to applications and offers topic-based discovery and data-processing services. Similar to GSN it tries to hide internals of sensors from the user but focuses on maintaining quality of service of data streams in the presence of disconnections while GSN is more targeted at flexible configurations, general abstractions, and distributed query support.

HiFi [5] provides efficient, hierarchical data stream query processing to acquire, filter, and aggregate data from multiple devices in a static environment while GSN takes a peer-to-peer perspective assuming a dynamic environment and allowing any node to be a data source, data sink, or data aggregator.

IrisNet [6] proposes a two-tier architecture consisting of sensing agents (SA) which collect

and pre-process sensor data and organizing agents (OA) which store sensor data in a hierarchical, distributed XML database. This database is modeled after the design of the Internet DNS and supports XPath queries. In contrast to that, GSN follows a symmetric peer-to-peer approach as already mentioned and supports relational queries using SQL.

Rooney et al. [8] propose so-called EdgeServers to integrate sensor networks into enterprise networks. EdgeServers filter and aggregate raw sensor data (using application specific code) to reduce the amount of data forwarded to application servers. The system uses publish/-subscribe style communication and also includes specialized protocols for the integration of sensor networks. While GSN provides a general-purpose infrastructure for sensor network deployment and distributed query processing, the EdgeServer system targets enterprise networks with application-based customization to reduce sensor data traffic in closed environments.

Besides these architectures, a large number of systems for query processing in sensor networks exist. Aurora [4] (Brandeis University, Braun University, MIT), STREAM [2] (Stanford), TelegraphCQ [3] (UC Berkeley), and Cougar [12] (Cornell) have already been discussed and related to GSN in Section 2.4.

In the Medusa distributed stream-processing system [13], Aurora is being used as the processing engine on each of the participating nodes. Medusa takes Aurora queries and distributes them across multiple nodes and particularly focuses on load management using economic principles and high availability issues. The Borealis stream processing engine [1] is based on the work in Medusa and Aurora and supports dynamic query modification, dynamic revision of query results, and flexible optimization. These systems focus on (distributed) query processing only, which is only one specific component of GSN, and focus on sensor heavy and server heavy application domains.

Additionally, several systems providing publish/subscribe-style query processing comparable to GSN exist, for example, [7].

C.3 Conclusions

The full potential of sensor technology will be unleashed through large-scale (up to global scale) data-oriented integration of sensor networks. To realize this vision of a “Sensor Internet” we suggest our Global Sensor Networks (GSN) middleware which enables fast and flexible deployment and interconnection of sensor networks. Through its virtual sensor abstraction which can abstract from arbitrary stream data sources and its powerful declarative specification and query tools, GSN provides simple and uniform access to the host of heterogeneous technologies. GSN offers zero-programming deployment and data-oriented integration of sensor networks and supports dynamic configuration and adaptation at runtime. Zero-programming deployment in conjunction with GSN’s plug-and-play detection and deployment feature provides a basic functionality to enable sensor mobility. GSN is implemented in Java and C/C++ and is available from SourceForge at http://gsn.sourcefourge.net/. The experimental evaluation of GSN demonstrates that the implementation is highly efficient, offers very good performance and throughput even under high loads and scales gracefully in the number of nodes, queries, and query complexity.