New distributed processing technology developed to efficiently collect desired data from big data streams

March 13 2012

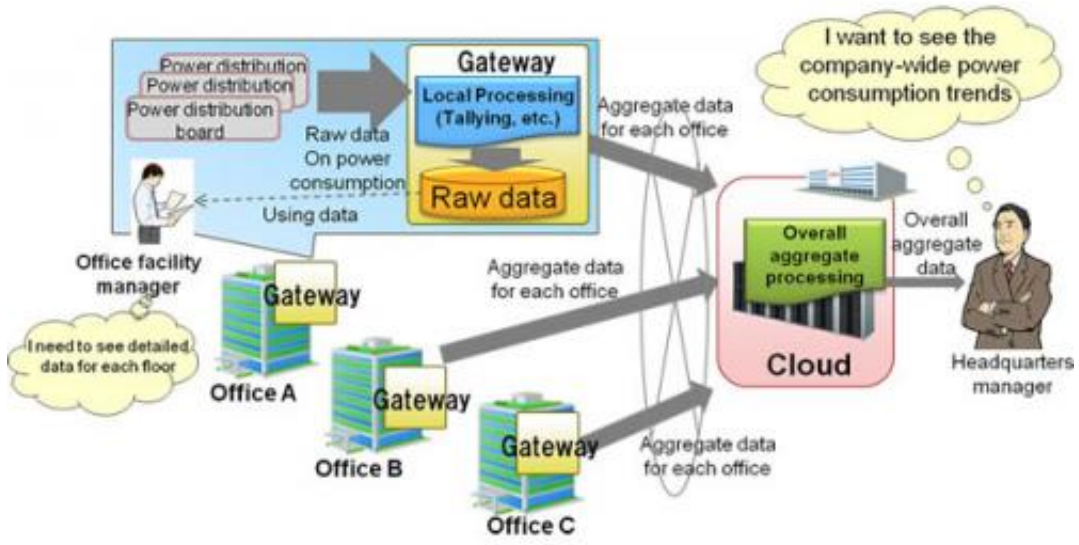

Figure 1. Distributing processing between gateways and cloud

Fujitsu Laboratories Limited today announced the world's first successful development of a distributed processing technology that efficiently collects enormous volumes of real-world data in the cloud through network-linked gateways. Services that collect and employ huge volumes of real-world data in the cloud—such as the location and health status of people and the status of different operations—are now expanding, and the increase in communications volumes associated with collecting this data poses a huge challenge. To address this problem,

Fujitsu Laboratories has developed an algorithm that takes a portion of the data that would otherwise be processed in the cloud and instead performs optimized distributed arrangement in a gateway. Using this technology, it is possible to efficiently collect in the cloud only the data required from the big data streams being processed through the gateway, enabling a 99% reduction in transmission traffic volumes.

As a result, huge volumes of real-world data can easily be used in the cloud while holding down communications costs. It is therefore hoped that this development will contribute to bringing about a human-centric intelligent society.

As devices and terminals have become more compact in recent years, and communications technology has become more sophisticated, we are starting to see the emergence of cloud services that, using a variety of machines and sensors that are connected to the network, collect and employ "big data." By collecting in the cloud the huge volumes of data being generated in the real world, it becomes possible to discern new insights that previously could not be detected, which can then be returned to society as new value. Examples would include the optimization of social infrastructure, such as power grid controls, or preventative maintenance of equipment through M2M (machine-to-machine) communications.

As more and more equipment is connected to the network, the increased volume of communications traffic is making it necessary to enhance cloud and communications network infrastructure, which in turn results in the major issue of higher communications costs. With the spread of smartphones in recent years, and with the next phase of connectivity encompassing not just people but also machines in a huge "network of things," the volume of communications traffic is expected to increase even more, giving rise to a strong need for technology that could reduce traffic volumes.

In many cases, the data generated from sensors or machines is not used as is, but first undergoes statistical or analytical processing before being put to use. Accordingly, by placing gateways that can perform some degree of processing nearer to where the data is generated, the data can, to the extent possible, be pre-processed, using filtering or statistical processing, near to the data source. Because just the processed data is then collected in the cloud, it is an effective approach to reducing communications traffic volumes.

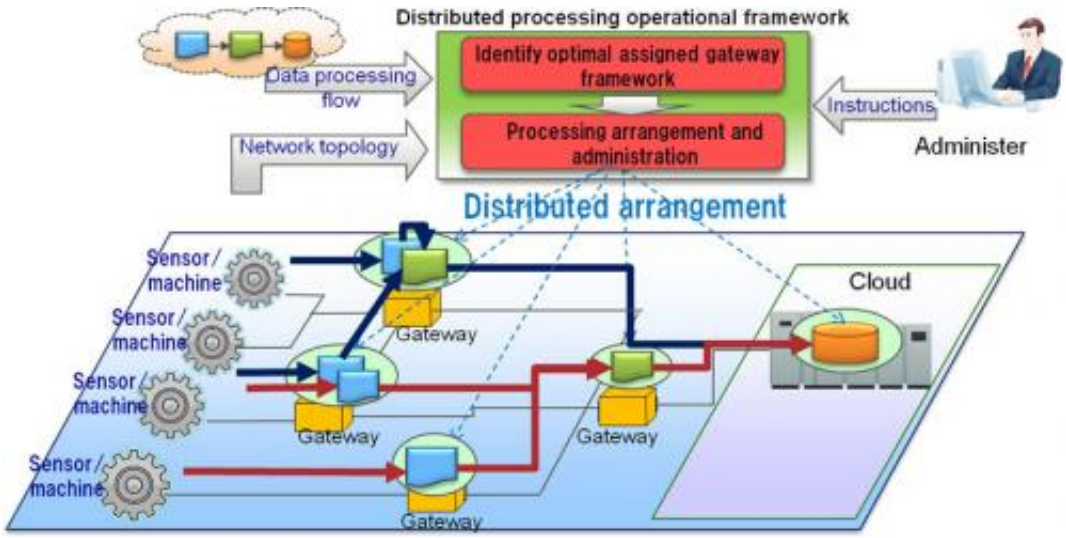

Figure 2. Overview of distributed processing operational framework

For example, as in figure 1 which illustrates electricity consumption, the raw data on electricity consumption for each office collected from power distribution boards or power supply taps is then converted into aggregate data for the company as a whole and displayed to managers in headquarters. By tallying this data in advance at gateways at each office and just sending the aggregate results to the cloud, the volume of data transmitted can be held down. Moreover, the raw data that was not sent to the cloud can later be compressed and sent to the cloud if truly

necessary, again reducing the volume of data compared to the situation in which raw data is sent individually.

Fujitsu Laboratories has used this approach to develop technology to process in gateways a portion of the data that would otherwise be processed in the cloud, thus reducing communication traffic volumes. The features of this technology are described below.

1. Distributed processing operational framework technology

Until now, distributed processing across multiple gateways of certain data that would otherwise be processed in the cloud, had required an administrator to judge what cloud processing at which gateway would effect a reduction in data volumes. Depending on the number of sensors, however, the number of gateways may change as gateways are added or eliminated. And because these changes, in turn, necessitate changes in the configuration of connections, the administrator would have to redesign the system each time, resulting in the problem of a higher administrative burden.

To address this problem, Fujitsu Laboratories developed an operational framework that automatically finds the gateways needed and assists in allocating and performing the processing program.

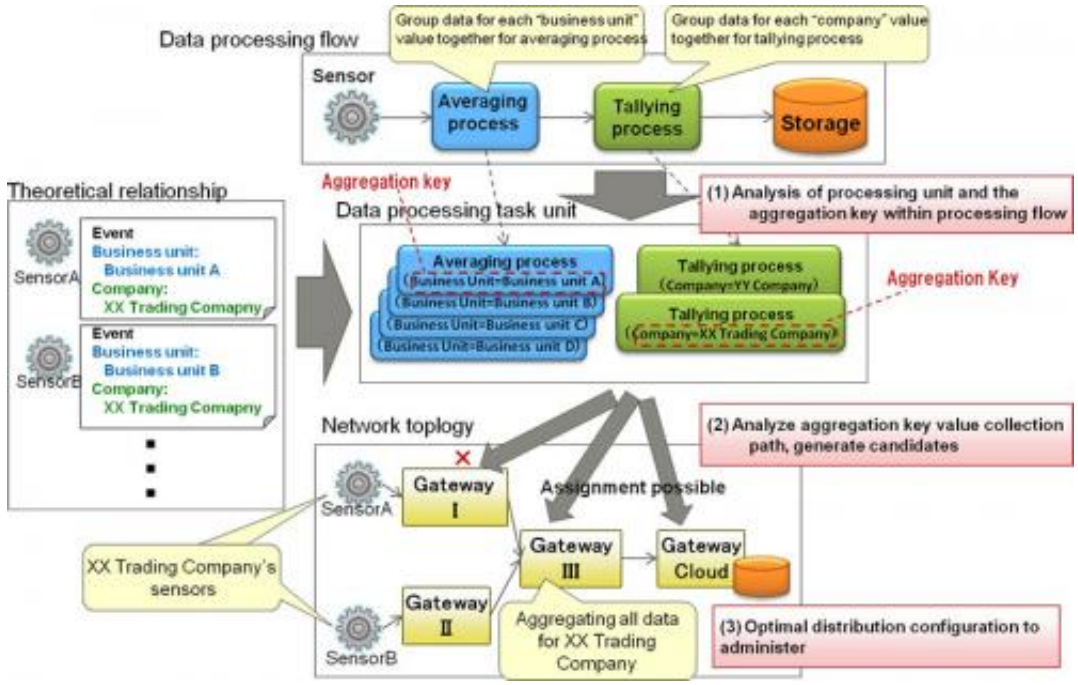

Figure 3. Flow of processes for optimal candidate extraction technology

The administrator defines the data processing program, whether it be tallying, filtering, or averaging, for example, as the flow of the data processing task units, and also defines the network topology that indicates the communications route from each sensor or machine to the gateway and then to the cloud. The framework, based on information on the data processing flow and the network topology, analyzes which gateways can perform each of the processing tasks within the overall data processing flow. It then generates the configuration of gateway assignments that minimizes the communications traffic volume. Even if the configuration of connections changes during operation, by redefining in the framework the new network topology, the gateway assignments can be reconfigured in accordance with the new network topology, enabling optimal distributed processing.

2. Aggregation key to identify distribution of the data

processing flow and optimal distributed arrangement

Averaging, tallying and other processing tasks that comprise the data processing flow are performed on aggregate data from multiple sensors. Therefore, the effective way of reducing communications traffic volumes is to perform the processing at the point where the data from these sensors is physically gathered together in one place. On the other hand, in such processing tasks as averaging or tallying, there are many cases in which the processing involves data that are theoretically aggregated, such as averages for each business unit or for the company as a whole. In such cases, in order to determine the assigned gateways, it is necessary to take into account not just the sensor data itself but also the theoretical meanings of the sensor data. Accordingly, Fujitsu Laboratories developed an aggregation key for grouping data for each form of processing, such as "by business unit" or "by company." It also developed a method for identifying the groups of sensors for each data processing task unit, based on the theoretical relationship between each sensor and such values as "business unit" or "company," and for efficiently identifying the optimal gateway for performing each data processing task unit.

In the example of figure 3, identification of assigned gateway candidates is performed in the flow described below.

a. Analysis of data processing flow

The aggregation key of the data processing flow is analyzed, and based on the theoretical relationship between the sensors and the values of the aggregation key, four averaging processes for the business units and the two tallying processes for the companies are extracted as processing units. Together with this, the sensors supplying the data for each data processing unit are identified.

b. Aggregation key value collection route analysis and selection of gateway candidates

By tracking the path taken in the network topology by the sensor data corresponding to the values of the aggregation key, it is determined that the tallying process for XX Trading Company cannot be performed at Gateway I or Gateway II; Gateway III and Gateway/Cloud are identified as assigned gateway candidates. Other assigned gateway candidates are similarly selected for each data processing task unit.

c. Generation of optimal arrangement configuration

From among the various configurations of assigned gateway candidates, the result that minimizes communications traffic volume is presented to the developer. When the developer selects the assigned gateways from among these candidates, the processing tasks are assigned to the gateways and performed.

With this technology, administrators using systems that handle big data can, without even having to be conscious of gateways, have processing tasks that had been performed in the cloud be automatically performed through distributed processing, with communications traffic volumes cut by 99%. As a result, cloud services employing huge volumes of real-world data can operate while holding down communications costs. For this reason, by offering a detailed service that can address situations with regard to people and environment, it is hoped that a contribution can be made to the realization of a human-centric intelligent society.

Fujitsu Laboratories will work on accelerating the speed of the algorithm for identifying assigned gateway candidates for the information gateway technology with aim of commercializing the technology in fiscal 2013.

Source: Fujitsu