Fermilab

Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

# The Resilient Smart Grid Workshop

## Network-based Data Service

October 16th, 2014

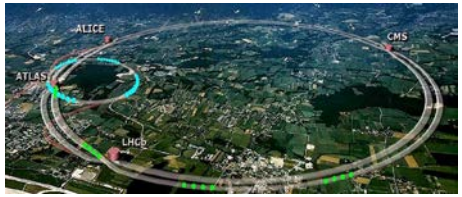
Jin Chang

# Agenda

- Fermilab Introduction
- Smart Grid Resilience Challenges
- Network-based Data Service (NDS) Introduction
- Network-based Data Service (NDS) Opportunities
- Next Steps

# Fermilab

- National Laboratory dedicated to HEP research
- More than 4,200 scientists worldwide use Fermilab and its particle accelerators, detectors and computers for their research, representing 34 countries, 120 U.S. institutions in 35 states.



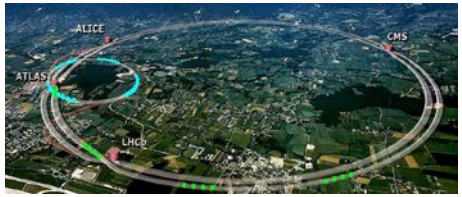

Study Higgs and search for new particles in high energy particle collisions

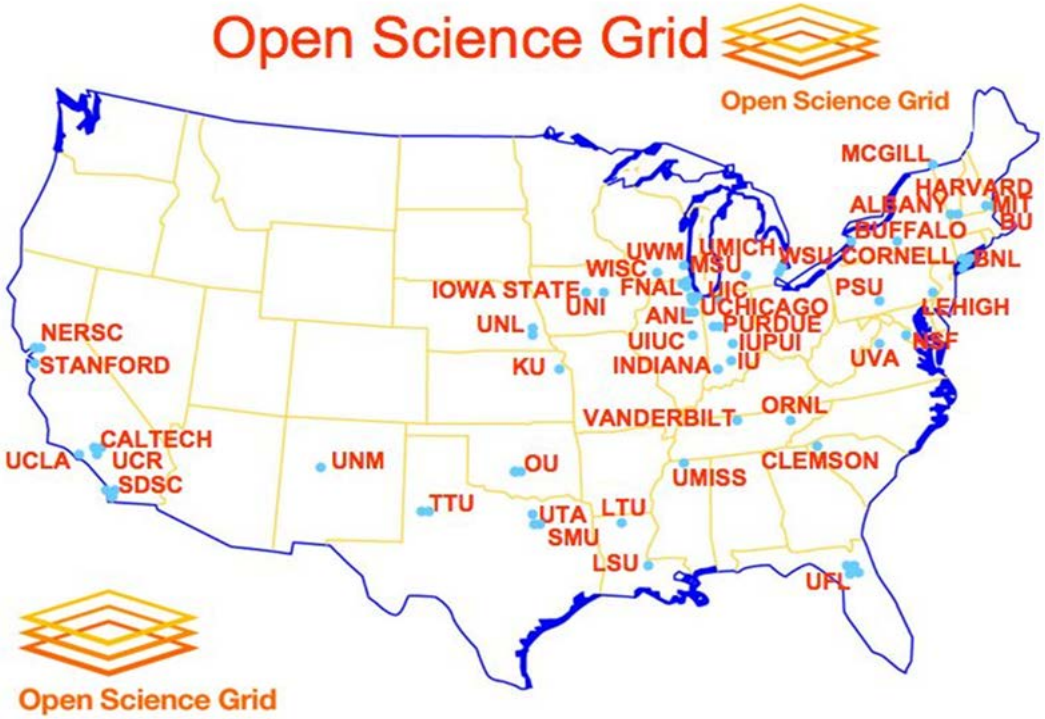
Study properties of neutrinos (oscillations, CP violation, mass)

Study muons with unprecedented precision (anomalous dipole moment, conversions)

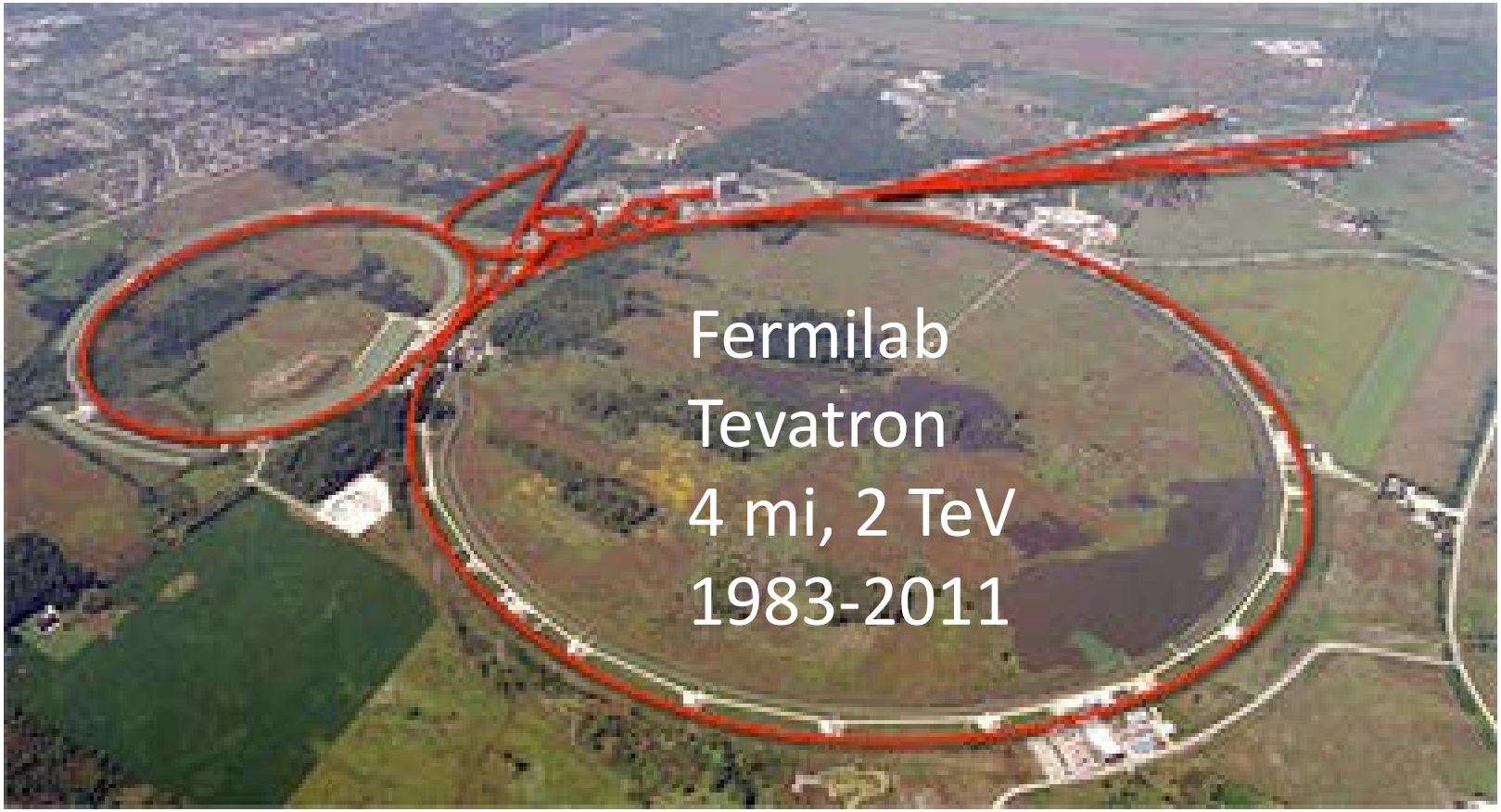
# Fermilab

- National Laboratory dedicated to HEP research
- More than 4,200 scientists worldwide use Fermilab and its particle accelerators, detectors and computers for their research, representing 34 countries, 120 U.S. institutions in 35 states.

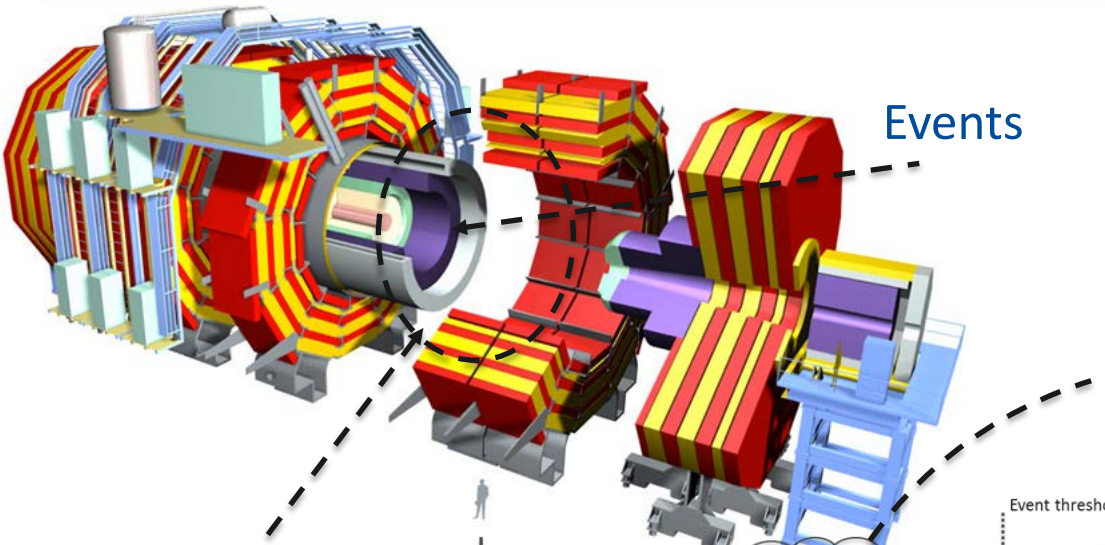




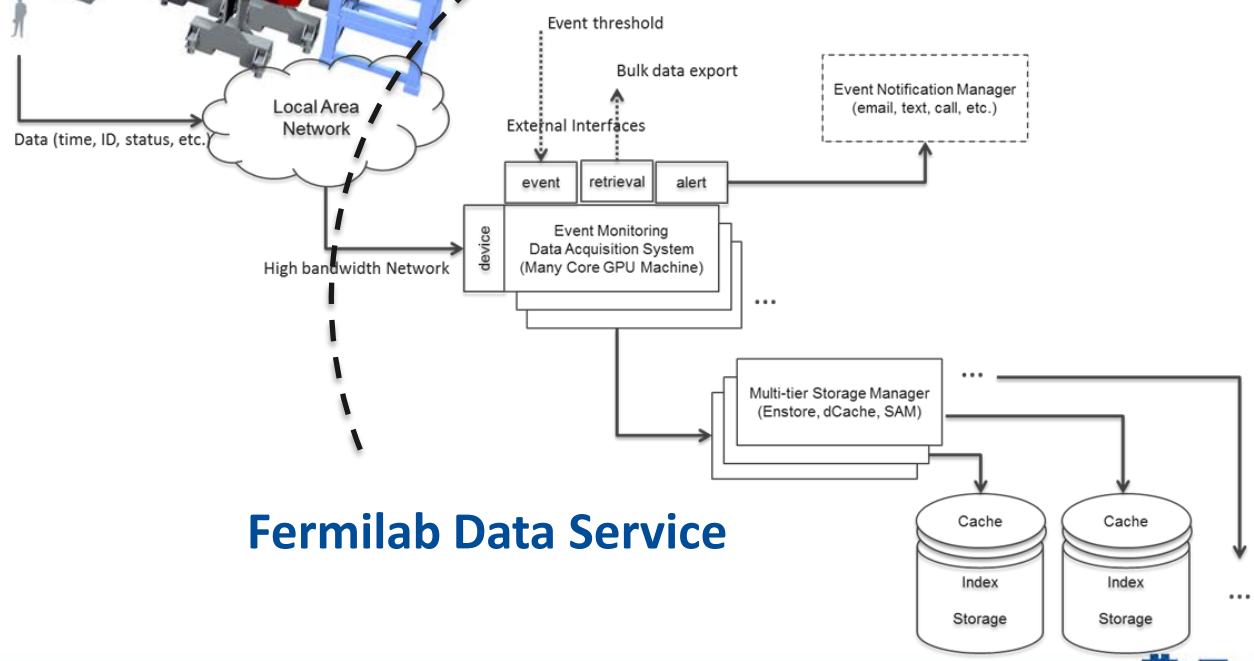
# High Energy Physics Colliders



# From High Energy Physics Experiments…




- Detectors built around collision point
- 100 Million individual measurements (channels) and sensors
- Trillions of events are recorded
- 20 PB of data per year




**Fermilab Data Service**

# Fermilab Computing Challenges

- Big Data
  - .75 Exabyte currently capable; 7 different tape robots
  - Hierarchical Mass Storage Systems with 30 PB disks cashe
- Networking
  - ~32,000 internal network ports
  - 112 Gb/s from 11 x 10 Gbps and 2 x 1 Gbps fiber links and additional 1 x 100 Gbps fiber link for research and development
- Computing Analysis
  - ~60,000 cores, high throughput / high performance computing
  - New technologies such as GPU's and Intel Phi
- Collaboration / Security
  - Researchers from ~40 different countries

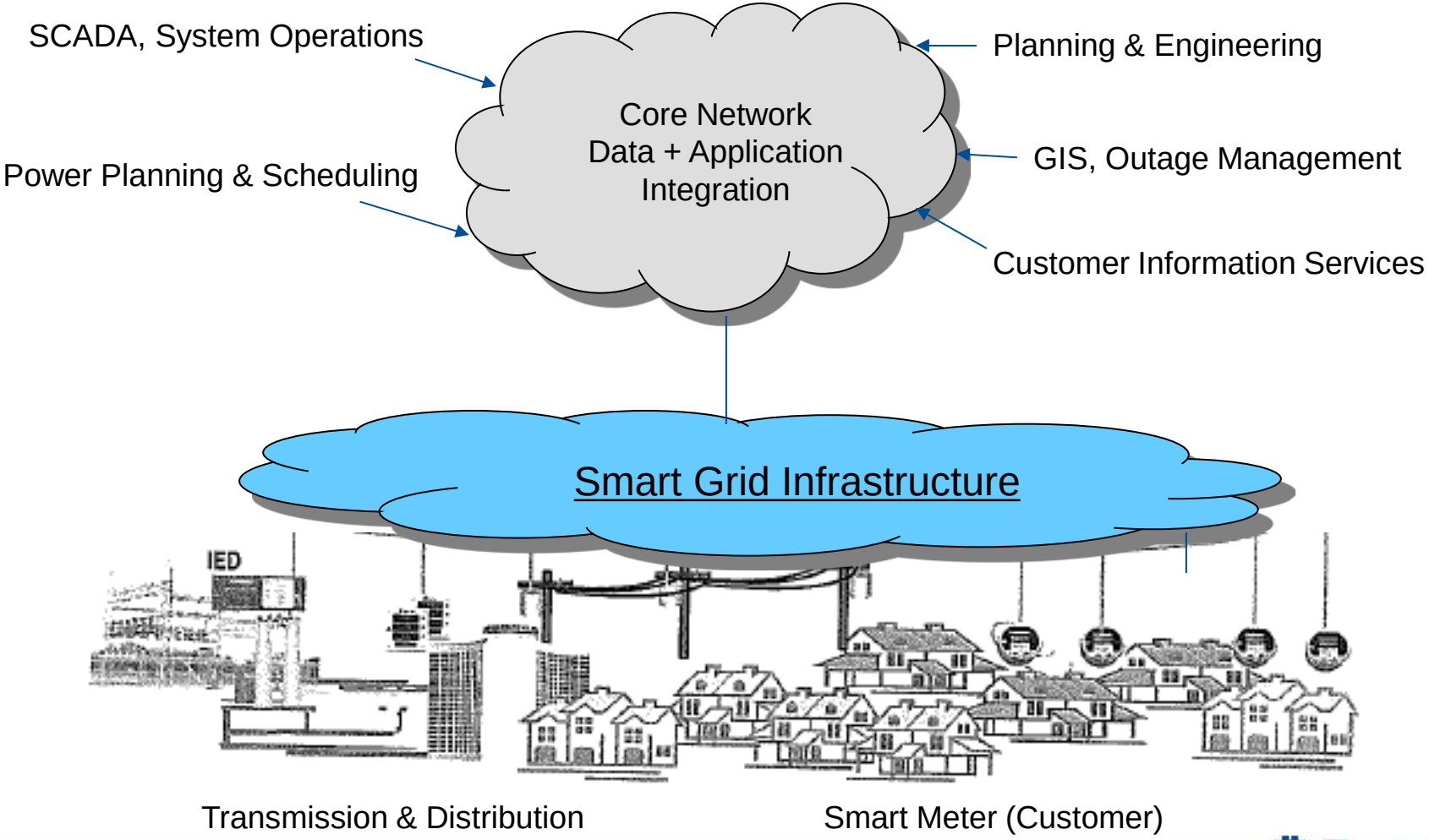
# Network-based Data Service (CEMAS)*

## for managing Internet of Everything (IoE) with focus on “smart city / smart grid”

**one of the 3 finalists for University of Chicago / Cisco EIR 2014 Innovation Challenge**

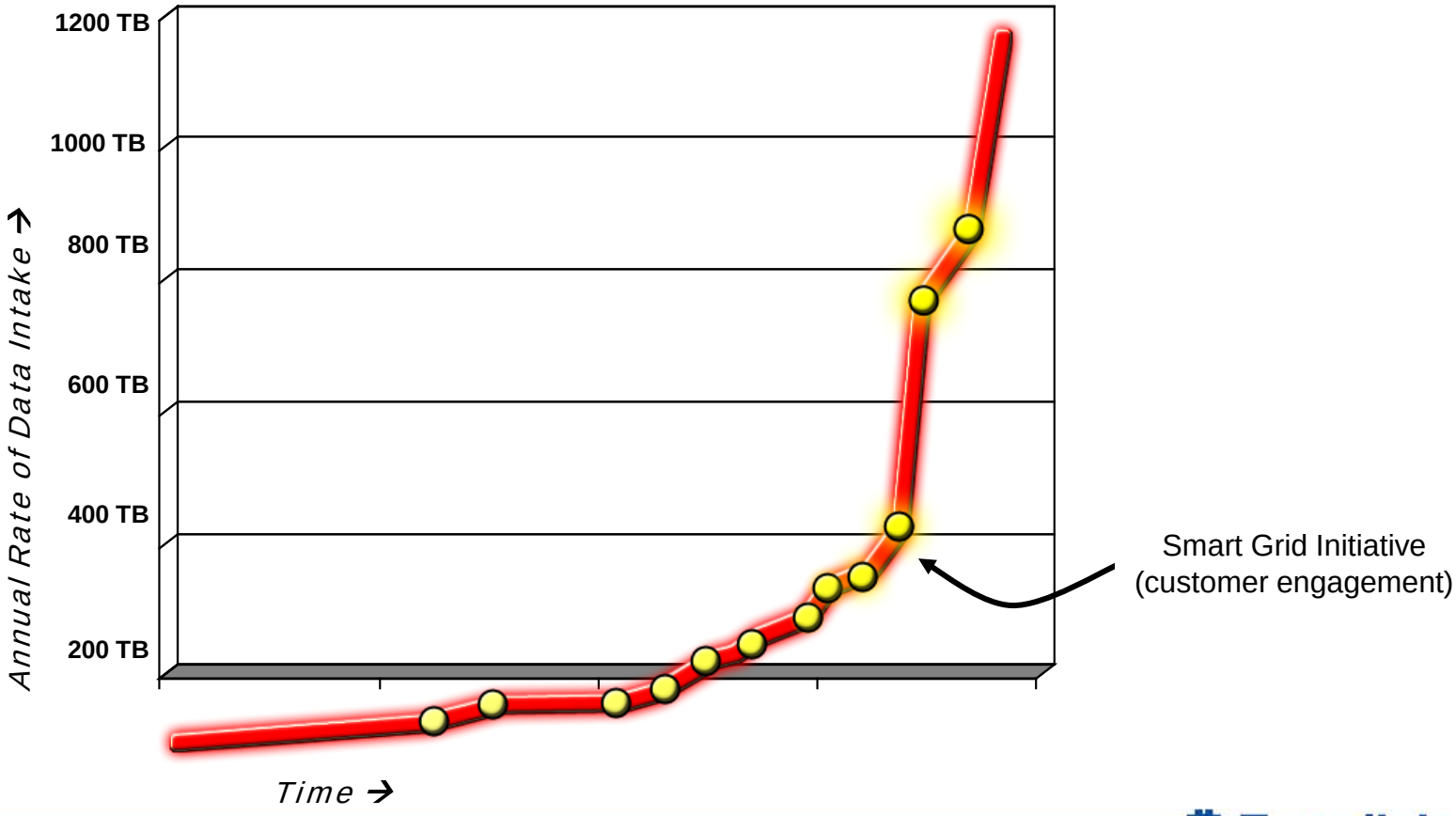
# Smart Grid Functional View

- Smart Grid technologies include many different domains that are influenced by many stakeholders, utilities, suppliers / vendors, regulators and customers.



# Smart Grid Data Challenge

- Smart Grid Initiative will experience significant increase in data volume
- Utility technologies and applications are not optimally architected
- Responding to important events / correlations becomes more challenging



# Smart Grid Conceptual Architecture

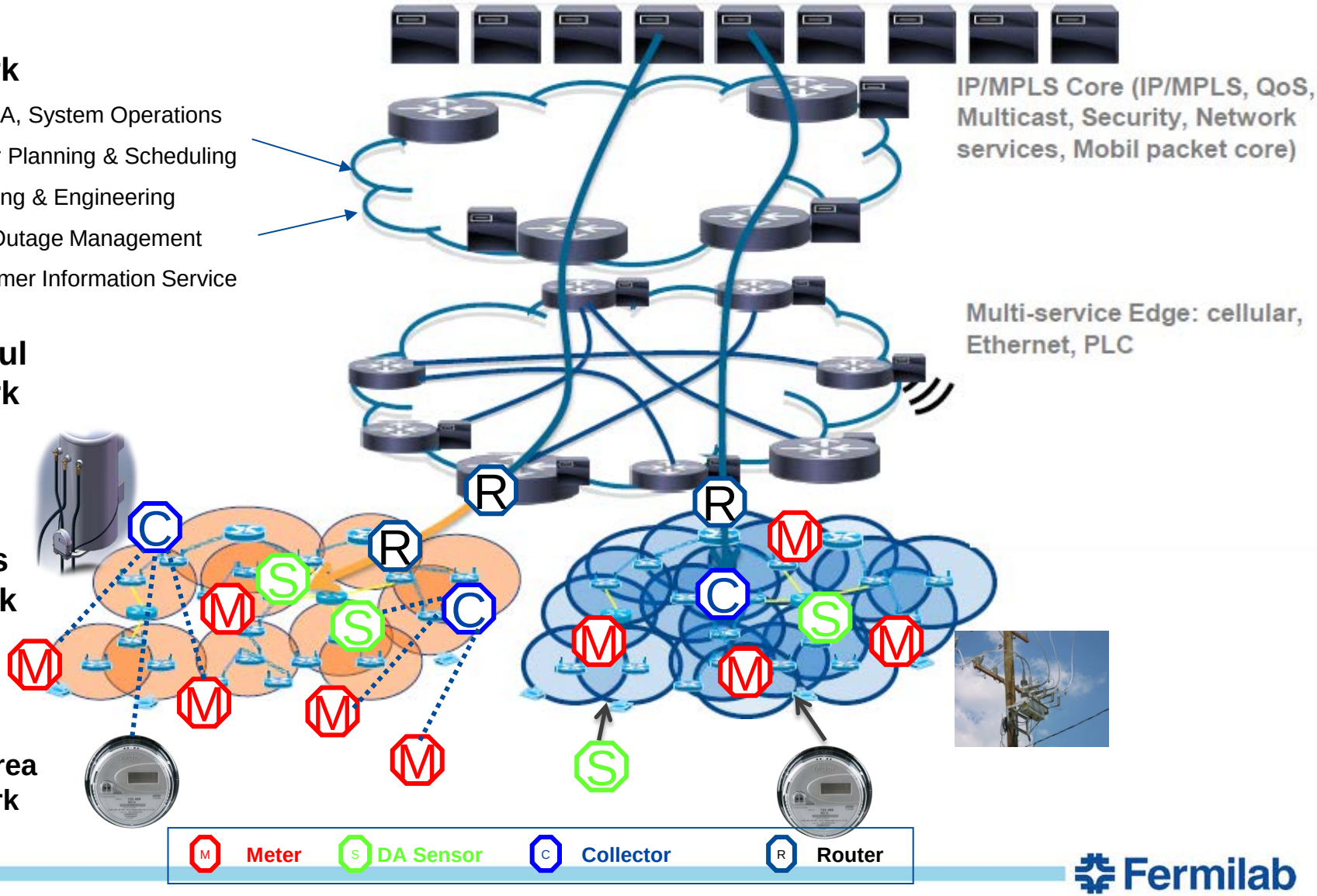
**Core Network**

- SCADA, System Operations
- Power Planning & Scheduling
- Planning & Engineering
- GIS, Outage Management
- Customer Information Service

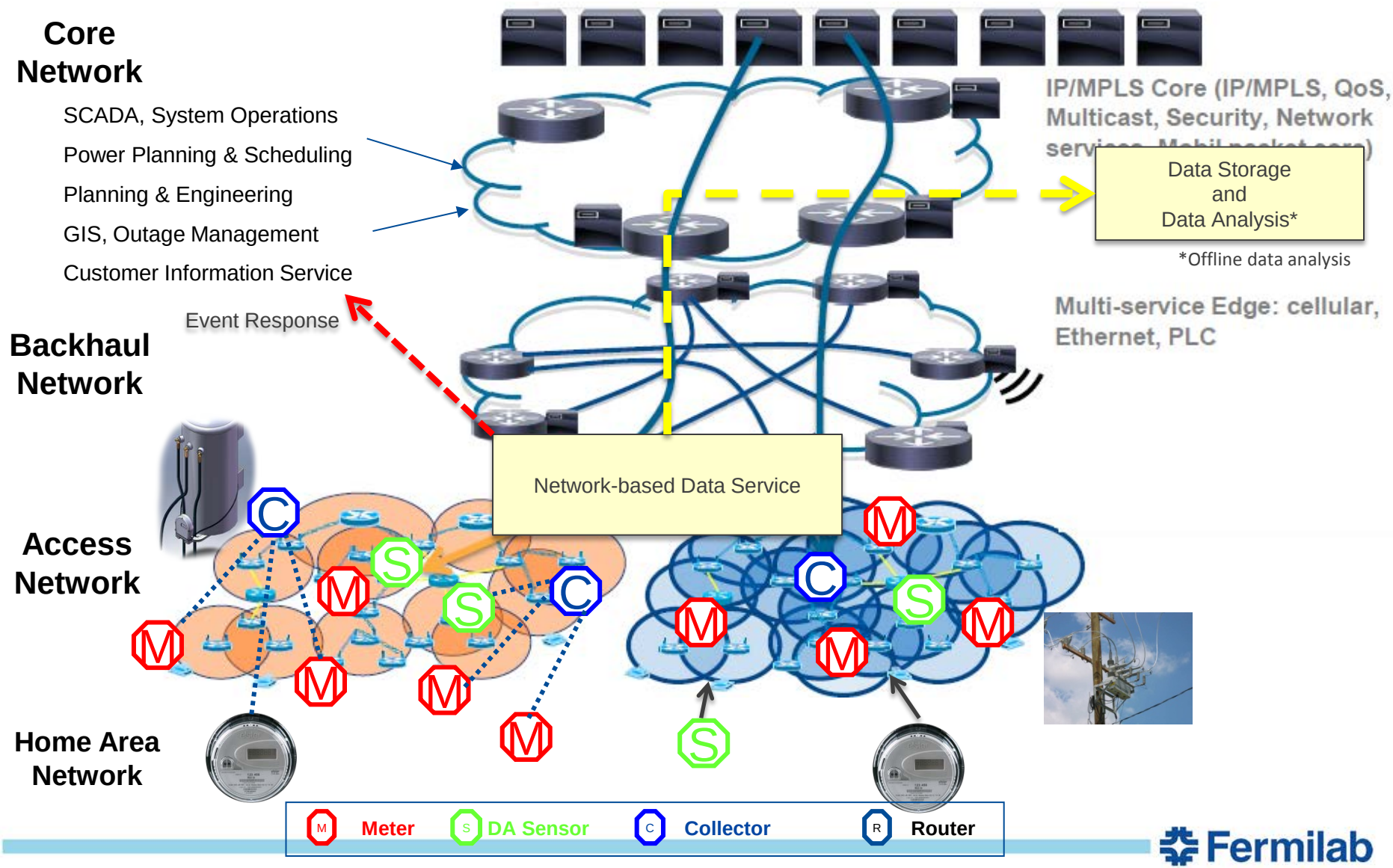
**Backhaul Network**

**Access Network**

**Home Area Network**



# Network-based Data Service Conceptual Architecture



# Network-based Data Service Advantages

- Data can be better managed and routed based on its signature and priority.
- Data quality of the devices can be evaluated in real time allowing early detection regarding any events.
- Sensors and devices can simply connect to the network without being tied to the implementation details of data management.
- The data acquisition and filtering rules can be managed centrally and the vendors can focus more on the reliability of grid devices instead of the infrastructure required for data management.
- Security policies can be applied uniformly and consistently at the network data service level for all the devices.

- Promote additional innovation and various services at the network level – somewhat similar to the Independent Producers and the *Paramount* Case (1949 )

# Network-based Data Service Challenges

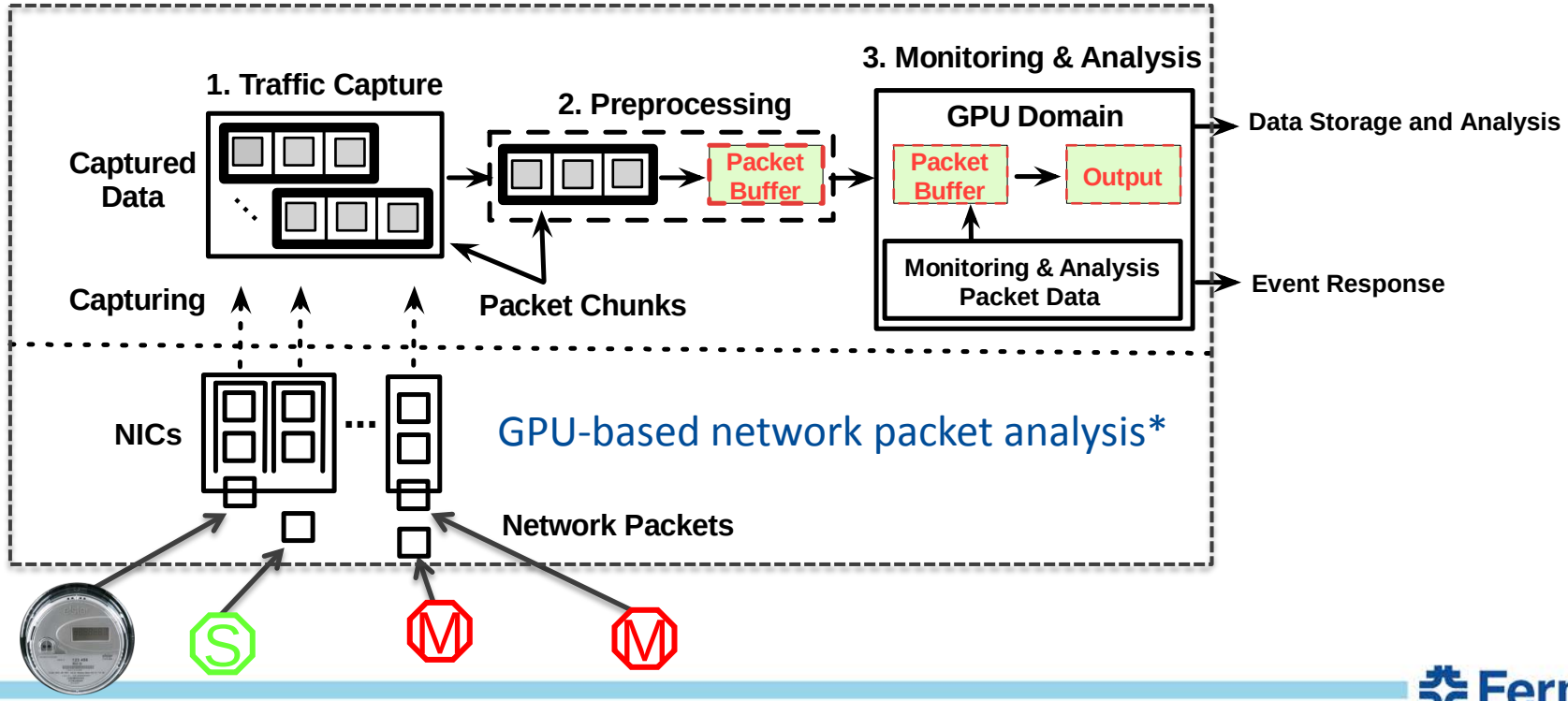
- Communication
  - Protocol differences (e.g., TCP/UDP)
  - Method variances
    - Push, Request / Response, Subscribe / Notify, Publish / Subscribe
  - High throughput / performance processing without packet loss

- Data Structure
  - Format (e.g., SGML, HTML, XML, etc.)
  - Storage and access (SQL, NoSQL, File/Meta Attribute, etc.)
  - Data mining and analytics

- Security

# Communication – Data Capture

- Ability to bring intelligence to network in order to tame the data stream.
  - 10 Gb/s network traffic monitoring; > 17 times faster than Xeon CPU
    http://www.networkworld.com/article/2172113/computers/super-computing-13--gpus-would-make-terrific-network-monitors.html
  - Performance benchmark demonstrating 30 times faster than Intel Xeon CPU
    http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6829552
- Part of the network infrastructure and no need for additional environment
- Potential to further increase the throughput (up to 100Gb/s) and high performance (up to 120 times faster than Intel Xeon CPU) via clustering & load balancing



# Data Storage Options

- However, these storage options require their own schema level conceptualization and translation of the data, hence limiting flexibility and resulting in delayed integration of various smart grid devices and sensors

http://www.datasciencecentral.com/profiles/blogs/row-vs-columnar-vs-nosql-databases

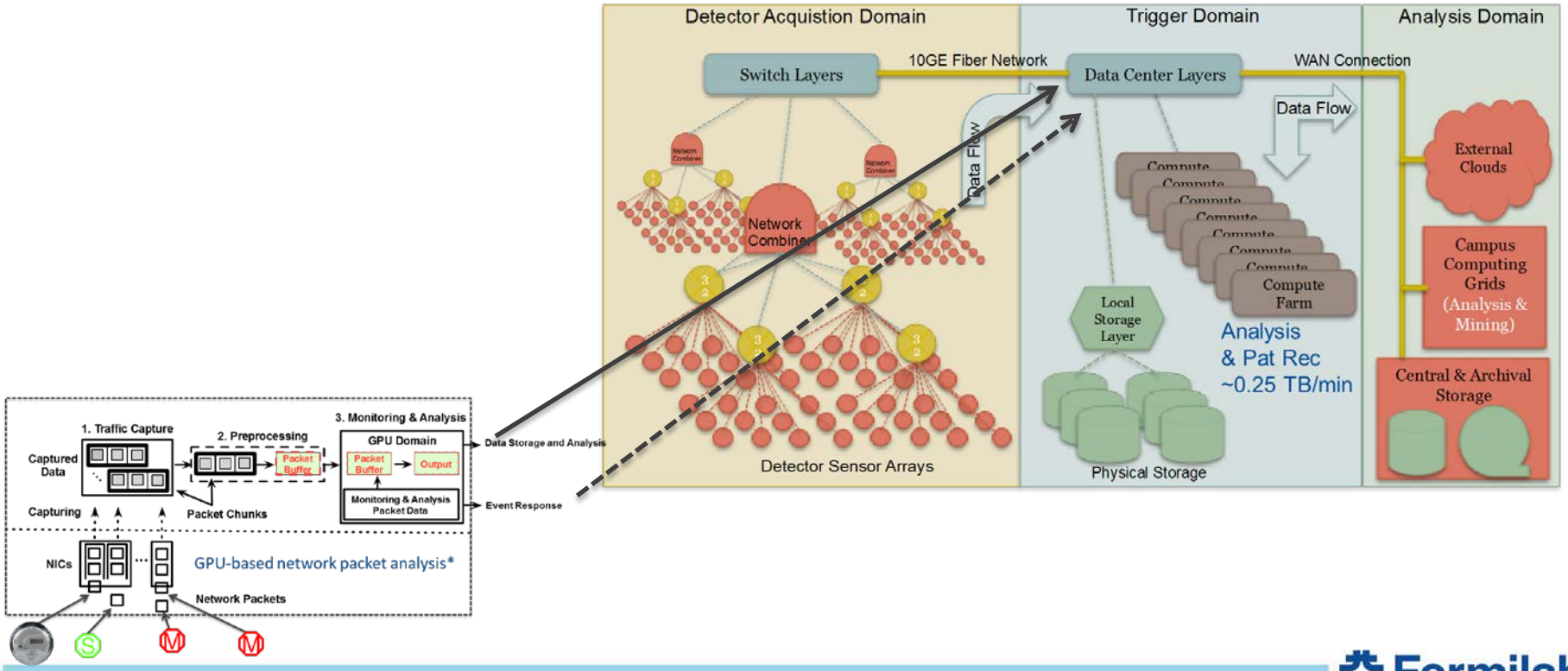
| | Row-Based | Columnar | NoSQL—Key Value Store | NoSQL—Document Store | NoSQL—Column Store |
|---|---|---|---|---|---|
| Basic Description | Data structured in rows | Data is vertically striped and stored in columns | Data stored usually in memory with some persistent backup | Persistent storage for unstructured or semi-structured data along with some SQL-like querying functionality | Very large data storage, MapReduc support |
| Common Use Cases | Transaction processing, interactive transactional applications | Historical data analysis, data warehousing, business intelligence | Used as a cache for storing frequently requested data for a web app | Web apps or any app which needs better performance and scalability without having to define columns in an RDBMS | Real-time data logging as in finance or web analytics |
| Strengths | Capturing and inputting new records. Robust, proven technology. | Fast query support, especially for ad hoc queries on large datasets, compression | Scalability, very fast storage and retrieval of unstructured and partly structured data | Persistent store with scalability features such as sharding built in with and better query support than key-value stores | Very high throughput for Big Data, strong partitioning support, randor read-write access |
| Weaknesses | Scale issues—less suitable for queries, especially against large databases | Not suited for transactions; import and export speed; heavy computing resource utilization | Usually all data must fit into memory, no complex query capabilities | Lack of sophisticated query capabilities | Low-level API, inability to perform complex queries, high latency of response to queries |
| Typical Database Size Range | | Several GBs to 50 TB | Several GBs to several TBs | Few TBs to several PBs | Few TBs to several PBs |
| Key Players | MySQL, Oracle, SQL Sever, Sybase ASE | Infobright, Aster Data, Sybase IQ, Vertica, ParAccel | MemCached, Amazon S3, Redis, Voldemort | MongoDb, Couchdb, SimpleDb | HBase, Big Table, Cassandra |

# Data Storage based on Object Store with Meta Attributes

- On the other hand, the raw data format with data descriptors such as time, location of devices, and duration offer greater flexibility to accommodate any differences among smart grid devices and vendors.
- It is expected that any performance deficiency associated with raw data format can be mitigated cheaply with intelligent data caching and additional back end processors.
- Compromised flexibility from structured data formats is more difficult to recover from.

# Big Data Management (SAM) Architecture

- Flexible object data storage based on sequential data access via meta data (SAM)
- Hierarchical data storage for managing big data (~200 peta bytes capable)
  - http://cd-docdb.fnal.gov/cgi-bin/RetrieveFile?docid=5085;filename=CHEP2013_SAM_IF.pdf
  - http://cd-docdb.fnal.gov/cgi-bin/RetrieveFile?docid=5082;filename=ifdh.pdf
- Existing Intelligent data caching system, enabling cost competitive to market leaders (e.g., better than 50% savings)
- Flexible data capture and acquisition for different types of cloud model (private vs. public)

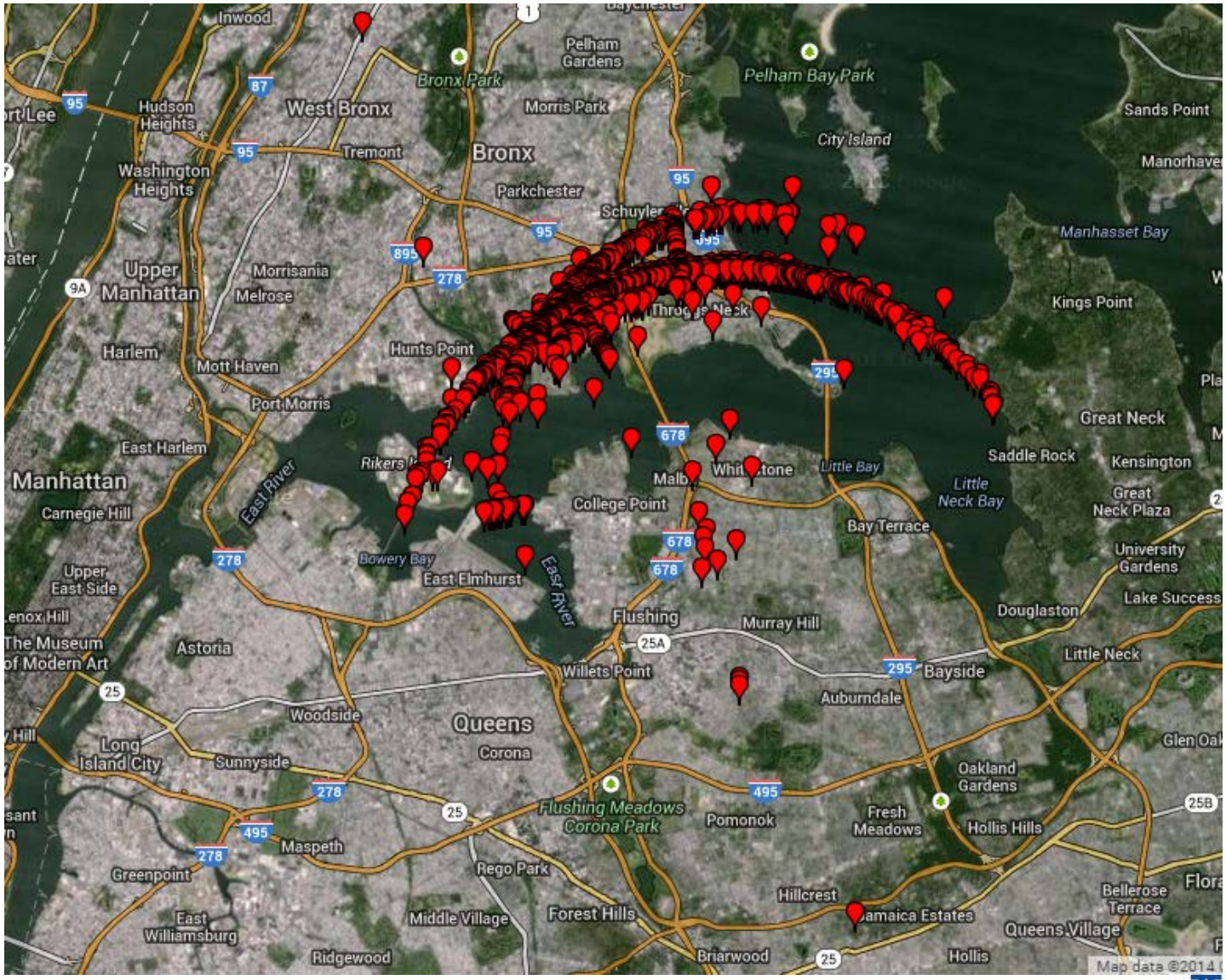


# Sensors / Devices Security Issues

- HP study* identified 70 percent of the most commonly used devices contain vulnerabilities.
  - Privacy concerns: 90 percent of tested devices collected at least one piece of personal information via the product itself, the cloud or its mobile application.
  - Insufficient authorization: 80 percent of IoT devices tested, including their cloud and mobile components, failed to require passwords of sufficient complexity
  - Lack of transport encryption: 70 percent of IoT devices analyzed did not encrypt communications to the internet and local network.
  - Insecure web interface: Six of the 10 devices evaluated raised security concerns with their user interfaces such as persistent XSS.
  - Inadequate software protection: 60 percent of devices did not use encryption when downloading software updates. Some downloads could even be intercepted, extracted and mounted as a file system in Linux where the software could be viewed or modified.

*July 29, 2014 HP release notes

# SCADA Security Incident



# Sensors / Devices Security Challenges

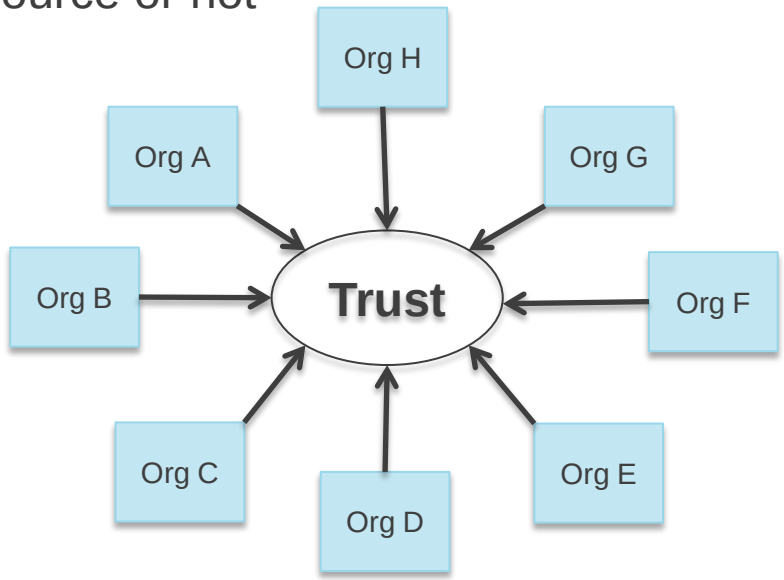
- Pre-shared Symmetric Keys (PSK) offer confidentiality and integrity.
- PSK has smaller computational and memory needs, suitable for constrained devices
- But, only works in local, closed environments. Key management for large number of nodes over the Internet, from different domains, randomly interacting with one another is **not scalable**

- Public Key Cryptography (PKC) does not have the key distribution problem, provides confidentiality, integrity, authentication
- **But computationally, it is beyond the capability of constrained devices.**
  - Most constrained devices do not have the RAM/ROM requirements and the computational power for PKC
- An alternative is Raw PKC which only has Public/Private key pairs without the certificate chain.
  - The challenge is where to register the Public Keys
- In general, PKC is difficult even for regular servers.
  - Maintaining the trust roots, the CRL downloads, keeping security libraries patched are all proven to be difficult even on servers with human admins

# Sensors / Devices Security based on Trust Federations

- Symmetric Keys combined with Trust Federations

- We already use them heavily for user-machine communication today
- InCommon Trust Federation in the US, connecting hundreds of universities, DOE Labs, companies
- Each organization has an Identity Provider vouching for its members identity
- An organization has a Resource Provider that decides whether access should be granted to the resource or not



# Next Steps

- We have an “unique” opportunity to promote broad, *underlying computing technologies* that are fundamental to “Society” and “Science”
  - Smart grid, smart city, IoT
  - HEP Experiments, biomedicine

- Through proper sponsorship from Government agencies and collaborating with representatives from the multiple national laboratories and other research organizations, the overall benefits will be significant
  - POC test, benchmark platform readily available

- Solicit participation on the upcoming “workshop” regarding *underlying computing technologies* – planned on February, 2015.
  - Government agencies and National Laboratories representatives
  - Technology companies, service organizations
  - Smart grid, smart city, SCADA and sensor/device vendors
  - Municipalities and Utilities – there are more than 400 different smart grid projects involving the various municipalities, utilities and other organizations*