Particle Physics Data Grid

Richard P. Mount
SLAC

DoE NGI PI Program PI Meeting
Washington DC, October 4, 1999

Particle Physics Data Grid

Universities, DoE Accelerator Labs, DoE Computer Science

- Particle Physics: a Network-Hungry Collaborative Application
  - Petabytes of compressed experimental data;
  - Nationwide and worldwide university-dominated collaborations analyze the data;
  - Close DoE-NSF collaboration on construction and operation of most experiments;
  - The PPDG lays the foundation for lifting the network constraint from particle-physics research.
- Short-Term Targets:
  - High-speed site-to-site replication of newly acquired particle-physics data (> 100 Mbytes/s);
  - Multi-site cached file-access to thousands of ~10 Gbyte files.

Collaborators:

California Institute of Technology	**Harvey B. Newman**, Julian J. Bunn, James C.T. Pool, Roy Williams
Argonne National Laboratory	**Ian Foster**, Steven Tuecke **Lawrence Price,** David Malon, Ed May
Berkeley Laboratory	**Stewart C. Loken**, Ian Hinchcliffe **Arie Shoshani,** Luis Bernardo, Henrik Nordberg
Brookhaven National Laboratory	**Bruce Gibbard,** Michael Bardash, Torre Wenaus
Fermi National Laboratory	**Victoria White**, Philip Demar, Donald Petravick **Matthias Kasemann,** Ruth Pordes
San Diego Supercomputer Center	**Margaret Simmons,** Reagan Moore,
Stanford Linear Accelerator Center	**Richard P. Mount,** Les Cottrell, Andrew Hanushevsky, David Millsom
Thomas Jefferson National Accelerator Facility	**Chip Watson,** Ian Bird
University of Wisconsin	**Miron Livny**

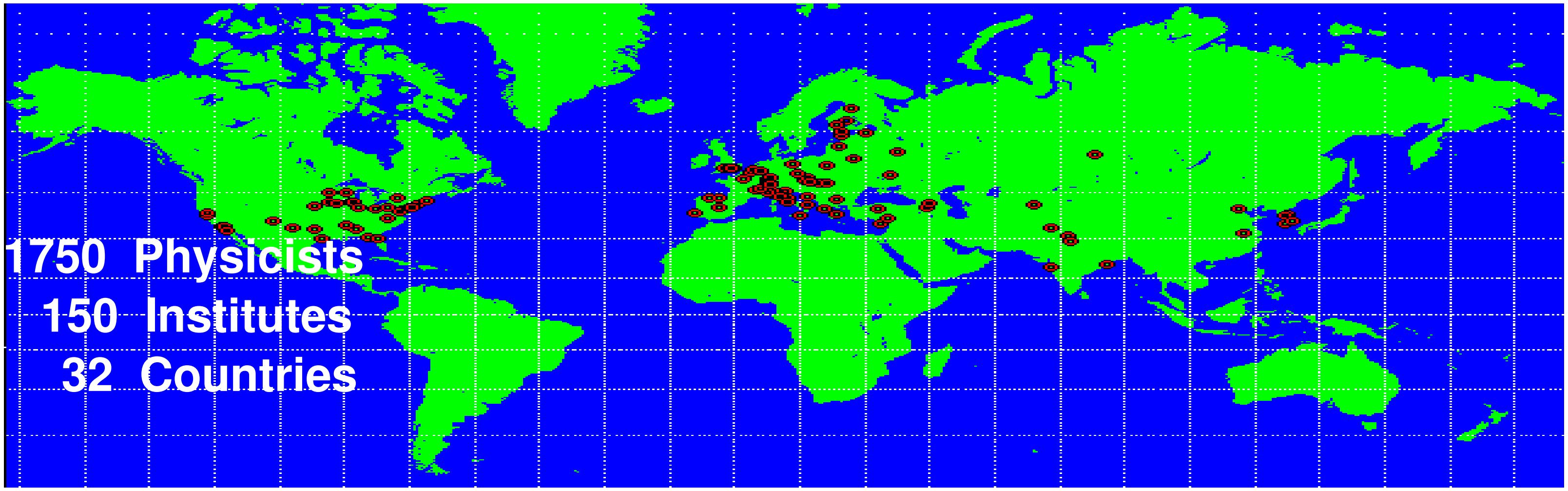
PPDG Collaborators

	Particle Physics	Accelerator Laboratory	Computer Science
ANL	X		X
LBNL	X		X
BNL	X	X	x
Caltech	X		X
Fermilab	X	X	x
Jefferson Lab	X	X	x
SLAC	X	X	x
SDSC			X
Wisconsin			X

Particle Physics Computing Challenges

- Geographical dispersion: of people and resources
- Complexity: the detector and the data
- Scale: Petabytes per year of data per experiment

Example: CMS Experiment

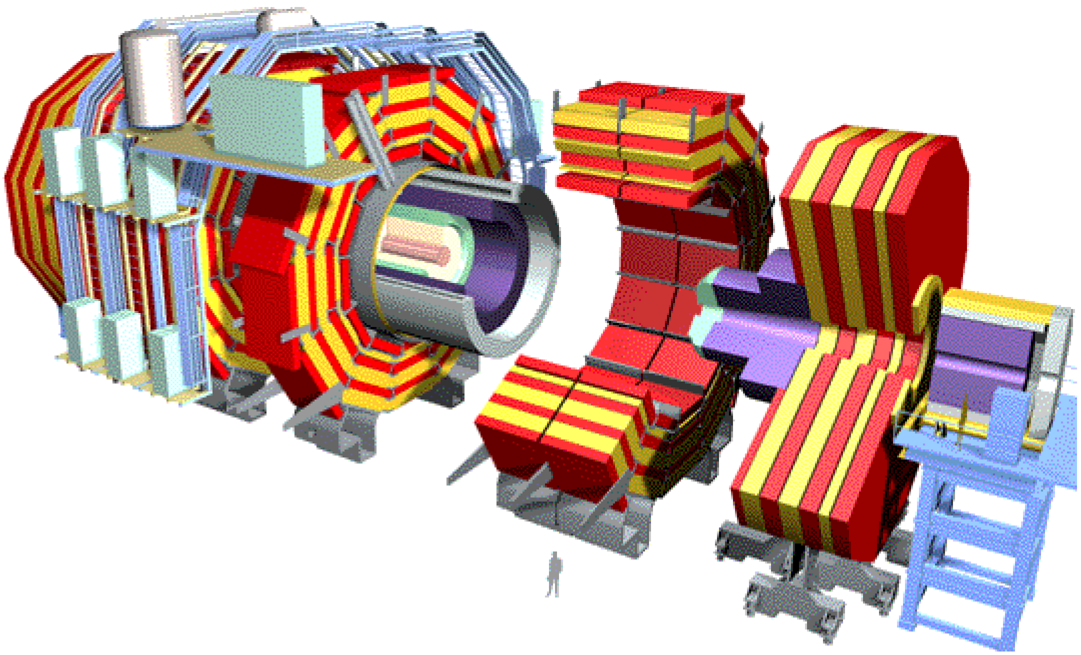


Major challenges associated with:

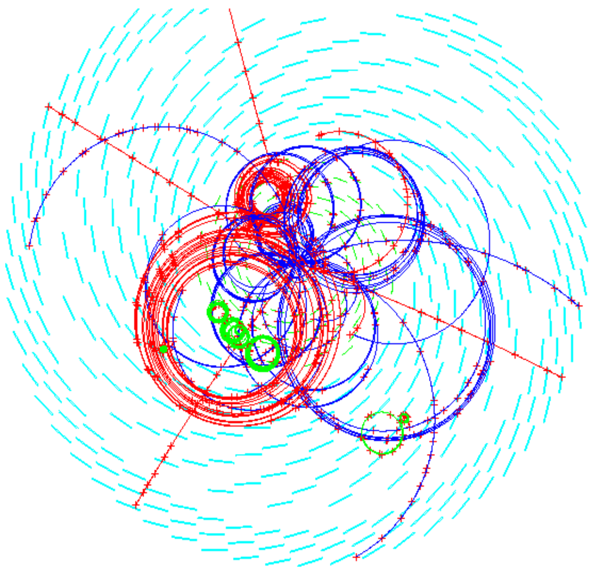
Communication and collaboration at a distance

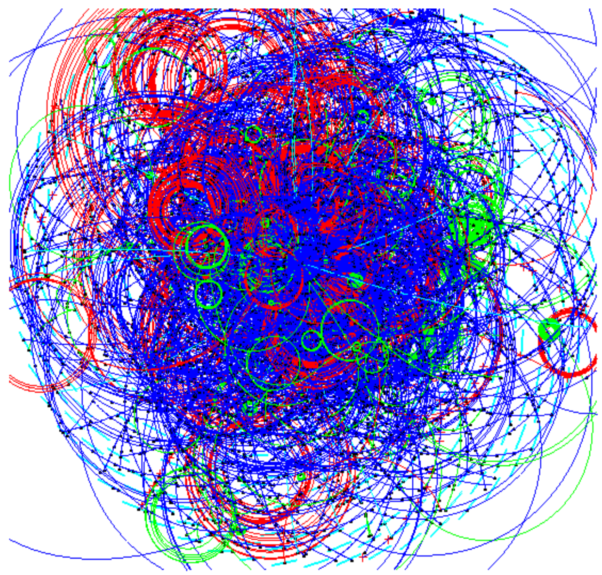
Distributed computing resources

Remote software development and physics analysis



Complex Detectors

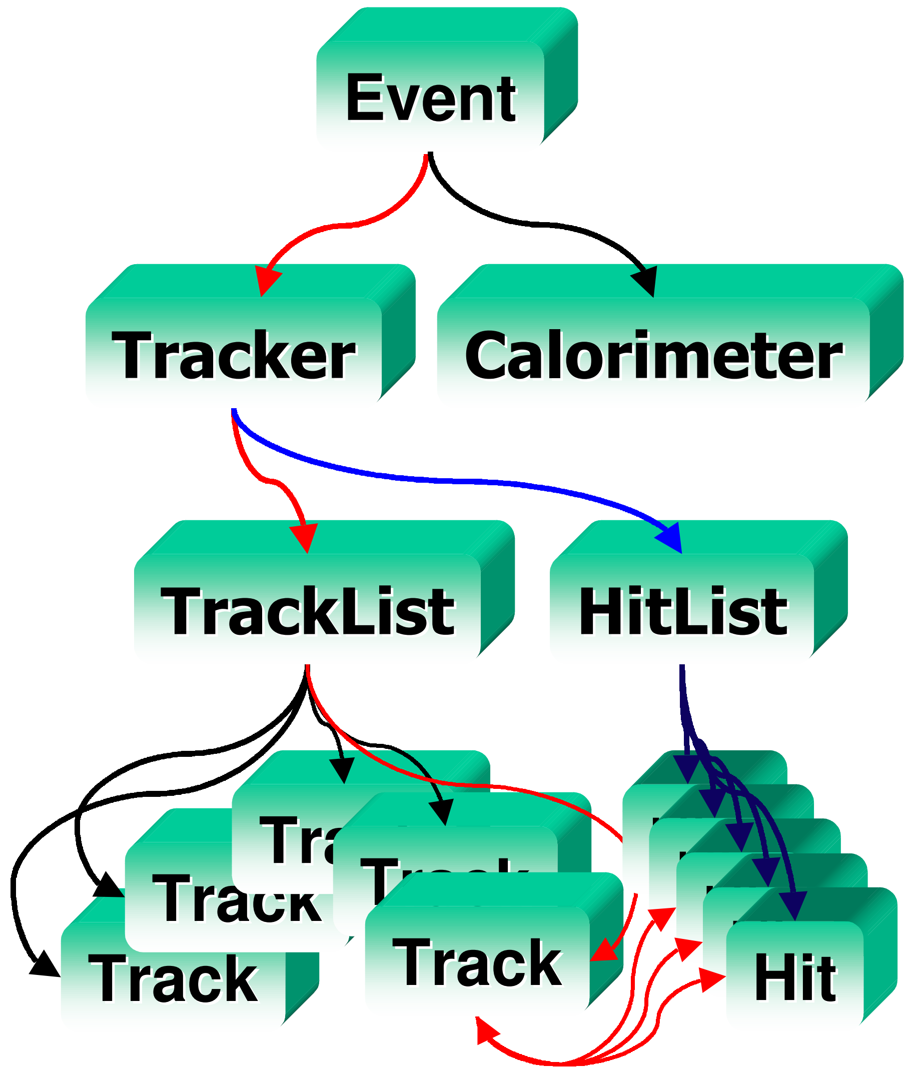




Complex Collision Products

Particle Physics Data Models

- **Particle physics data models are complex!**
  - **Rich hierarchy of hundreds of complex data types (classes)**
  - **Many relations between them**
  - **Different access patterns (Multiple Viewpoints)**

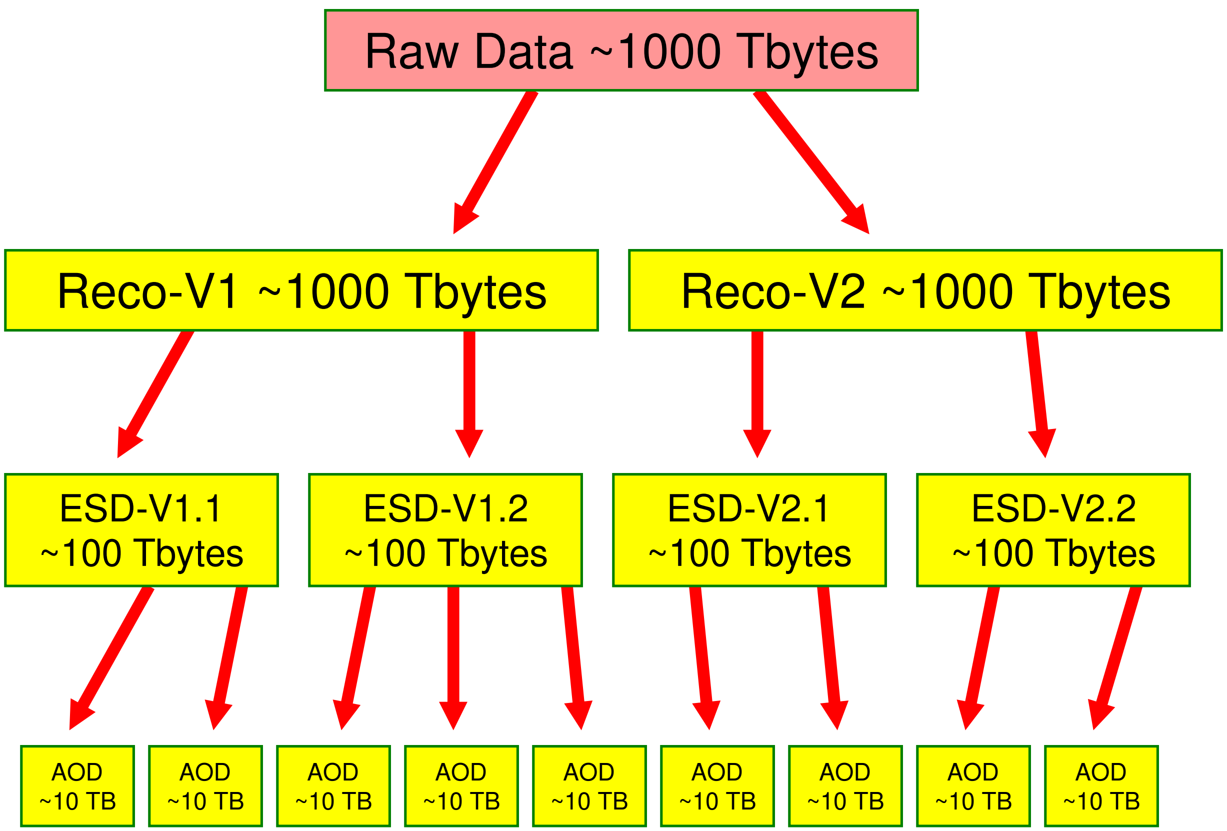


Data Volumes

- Quantum Physics yields predictions of probabilities;
- Understanding physics means measuring probabilities;
- Precise measurements of new physics require analysis of hundreds of millions of collisions (each recorded collision yields ~1Mbyte of compressed data)

Access Patterns

Typical particle physics experiment in 2000-2005:
On year of acquisition and analysis of data



Access Rates (aggregate, average)

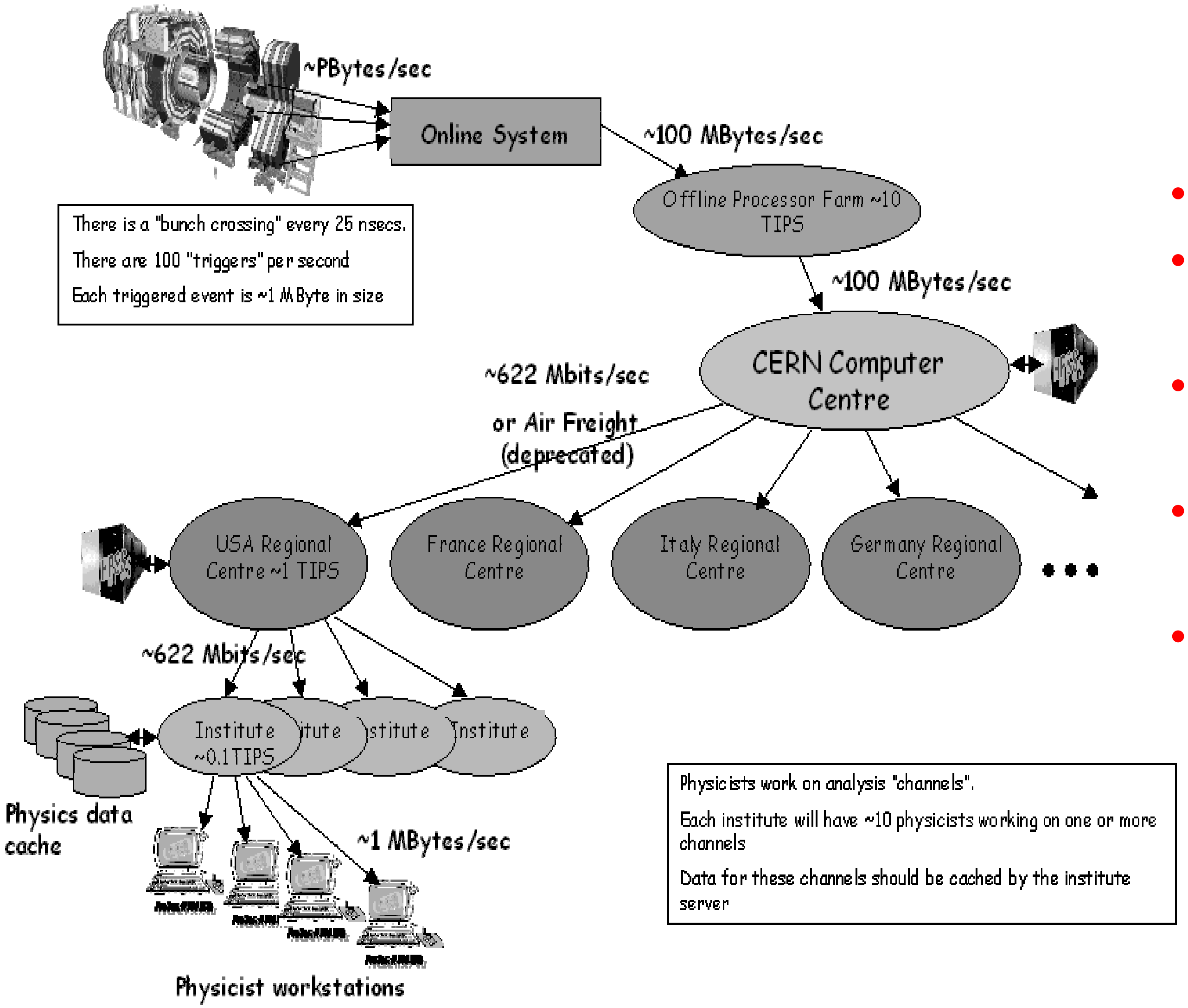
100 Mbytes/s (2-5 physicists)

1000 Mbytes/s (10-20 physicists)

2000 Mbytes/s (~100 physicists)

4000 Mbytes/s (~300 physicists)

Data Grid Hierarchy Regional Centers Concept



- **LHC Grid Hierarchy Example**
- **Tier0: CERN**
- **Tier1: National "Regional" Center**
- **Tier2: Regional Center**
- **Tier3: Institute Workgroup Server**
- **Tier4: Individual Desktop**
- **Total 5 Levels**

PPDG as an NGI Problem

PPDG Goals

The ability to query and partially retrieve hundreds of terabytes across Wide Area Networks within seconds,

Making effective data analysis from ten to one hundred US universities possible.

PPDG is taking advantage of NGI services in three areas:

– **Differentiated Services:** to allow particle-physics bulk data transport to coexist with interactive and real-time remote collaboration sessions, and other network traffic.

– **Distributed caching:** to allow for rapid data delivery in response to multiple "interleaved" requests

– **"Robustness": Matchmaking and Request/Resource co-scheduling:** to manage workflow and use computing and net resources efficiently; to achieve high throughput

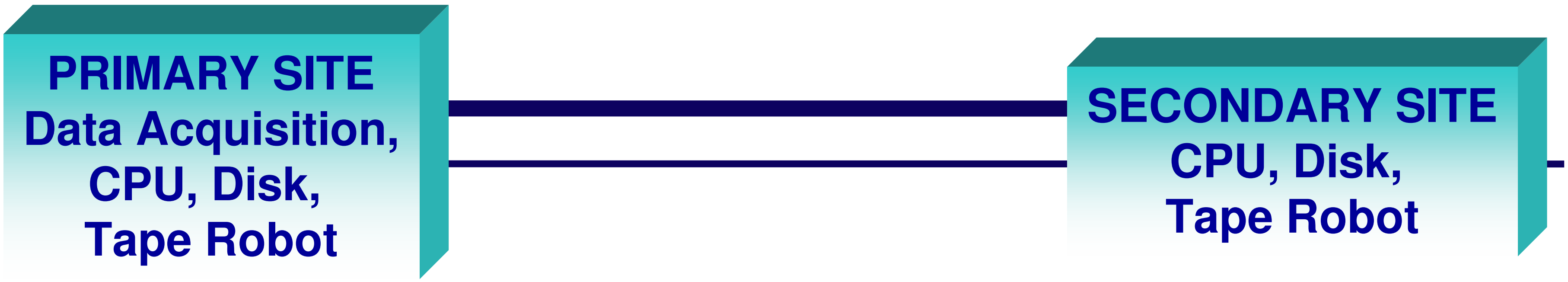
First Year PPDG Deliverables

Implement and Run two services in support of the major physics experiments at BNL, FNAL, JLAB, SLAC:

- **"High-Speed Site-to-Site File Replication Service";**
  Data replication up to 100 Mbytes/s
- **"Multi-Site Cached File Access Service":**
  Based on deployment of file-cataloging, and transparent cache-management and data movement middleware
- **First Year: Optimized cached read access to file in the range of 1-10 Gbytes, from a total data set of order One Petabyte**

Using middleware components already developed by the Proponents

PPDG Site-to-Site Replication Service



- Network Protocols Tuned for High Throughput
- Use of DiffServ **for**
  - (1) Predictable high priority delivery of high - bandwidth data streams
  - (2) Reliable background transfers
- Use of integrated instrumentation
  to detect/diagnose/correct problems in long-lived high speed transfers [NetLogger + DoE/NGI developments]
- Coordinated reservaton/allocation techniques
  for storage-to-storage performance

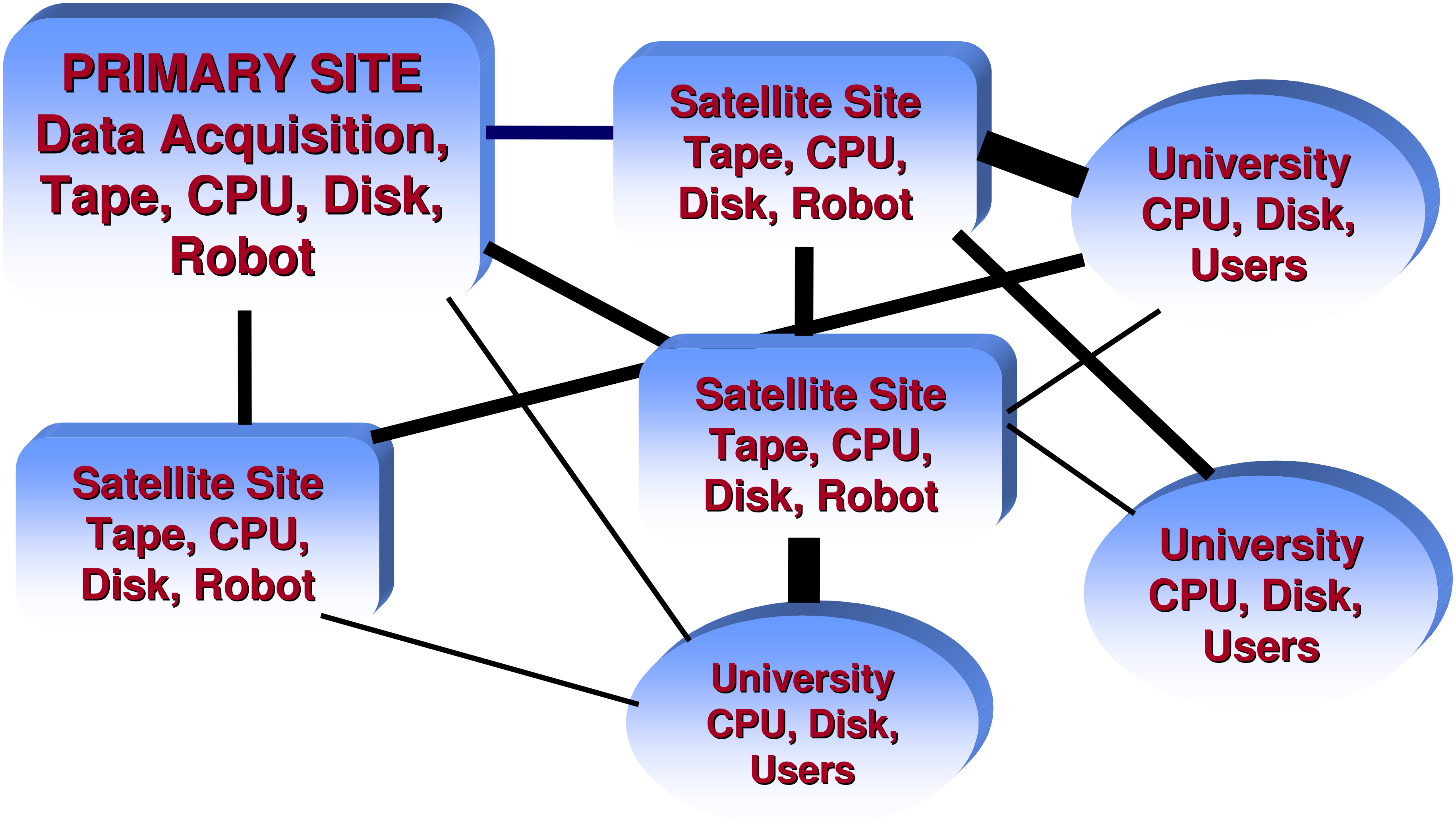
Typical HENP Primary Site ~Today (SLAC)

- 15 Tbytes disk cache
- 800 Tbytes robotic tape capacity
- 10,000 Specfp/Specint 95
- Tens of Gbit Ethernet connections
- Hundreds of 100 Mbit/s Ethernet connections
- Gigabit WAN access.

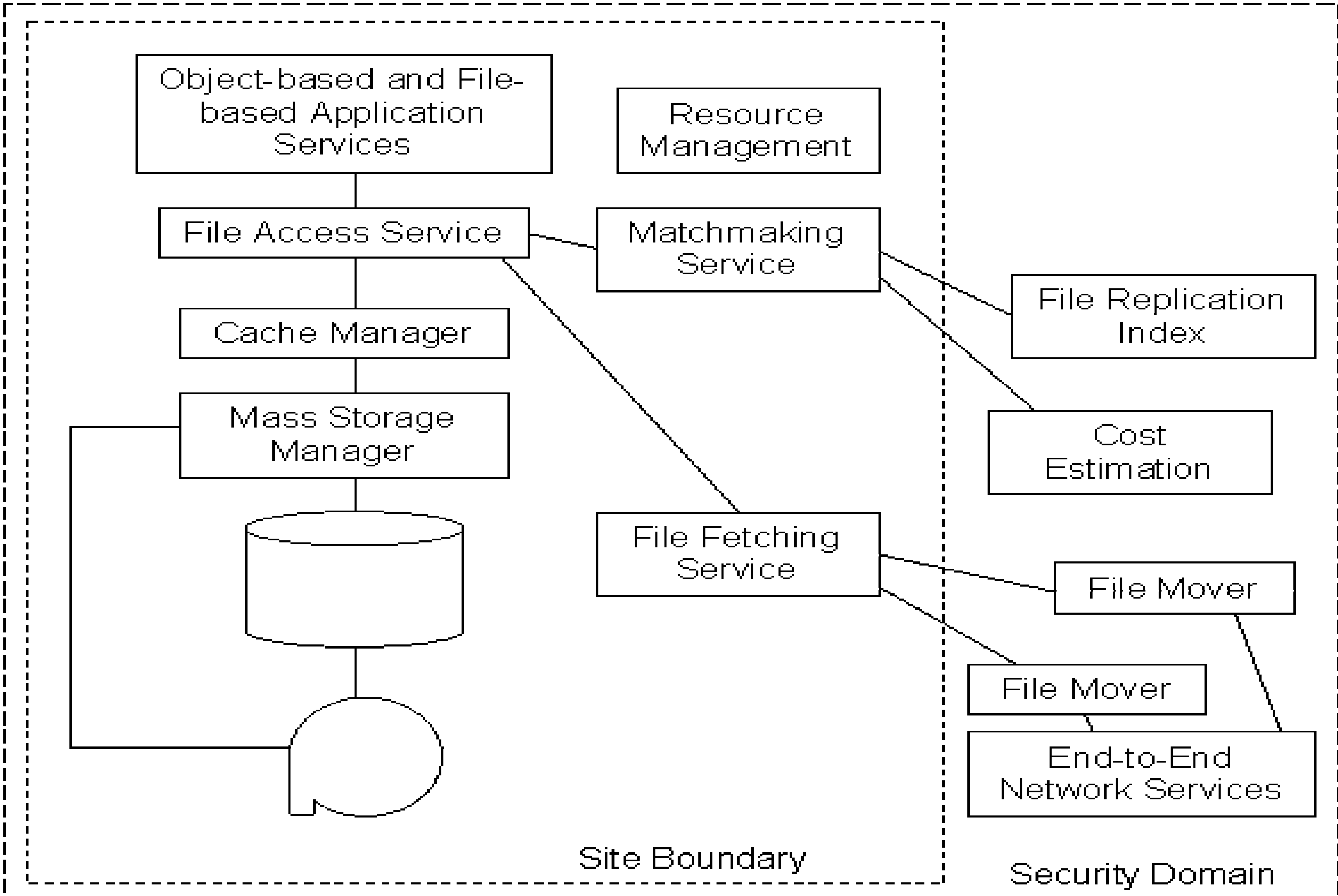
Data Center Resources Relevant to FY 1999 Program of Work

Site	CPU Gigaops/s	Mass Storage Management Software	Disk Cache Terabytes	Robotic Tape Storage Terabytes	Network Connections	Network Access Speeds
ANL	100	HPSS	>1	80	ESnet MREN	OC12 OC3-OC48
BNL	400	HPSS	20	600	ESnet	OC3
Caltech	100	HPSS	1.5	300	NTON CalREN-2 CalREN-2 ATM ESnet (direct)	OC12- (OC48) OC12 OC12 T1
Fermi Lab	100	Enstore HPSS	5	100	ESnet MREN	OC3 OC3
Jefferson Lab	80	OSM	3	300	ESnet	T3-(OC3)
LBNL	100	HPSS	1	50	ESnet CalREN-2 NTON	OC12 OC12 OC12- OC48
SDSC					CalREN-2 NTON ESnet	OC12 OC12- OC48 OC3
SLAC	300	HPSS	10	600	NTON ESnet	OC12- OC48 OC3
U. Wisconsin	~100				MREN	OC3

PPDG Multi-site Cached File Access System



PPDG Middleware Components



First Year PPDG "System" Components

Middleware Components (Initial Choice): See PPDG Proposal Page 15

Object and File-Based Application Services	**Objectivity/DB (SLAC enhanced)** **GC Query Object, Event Iterator, Query Monitor** **FNAL SAM System**
Resource Management	**Start with Human Intervention**

(but begin to deploy resource discovery & mgmnt tools)

File Access Service	**Components of OOFS (SLAC)**
Cache Manager	**GC Cache Manager (LBNL)**
Mass Storage Manager	**HPSS, Enstore, OSM (Site-dependent)**
Matchmaking Service	**Condor (U. Wisconsin)**
File Replication Index	**MCAT (SDSC)**
Transfer Cost Estimation Service	**Globus (ANL)**
File Fetching Service	**Components of OOFS**
File Movers(s)	**SRB (SDSC); Site specific**
End-to-end Network Services	**Globus tools for QoS reservation**
Security and authentication	**Globus (ANL)**

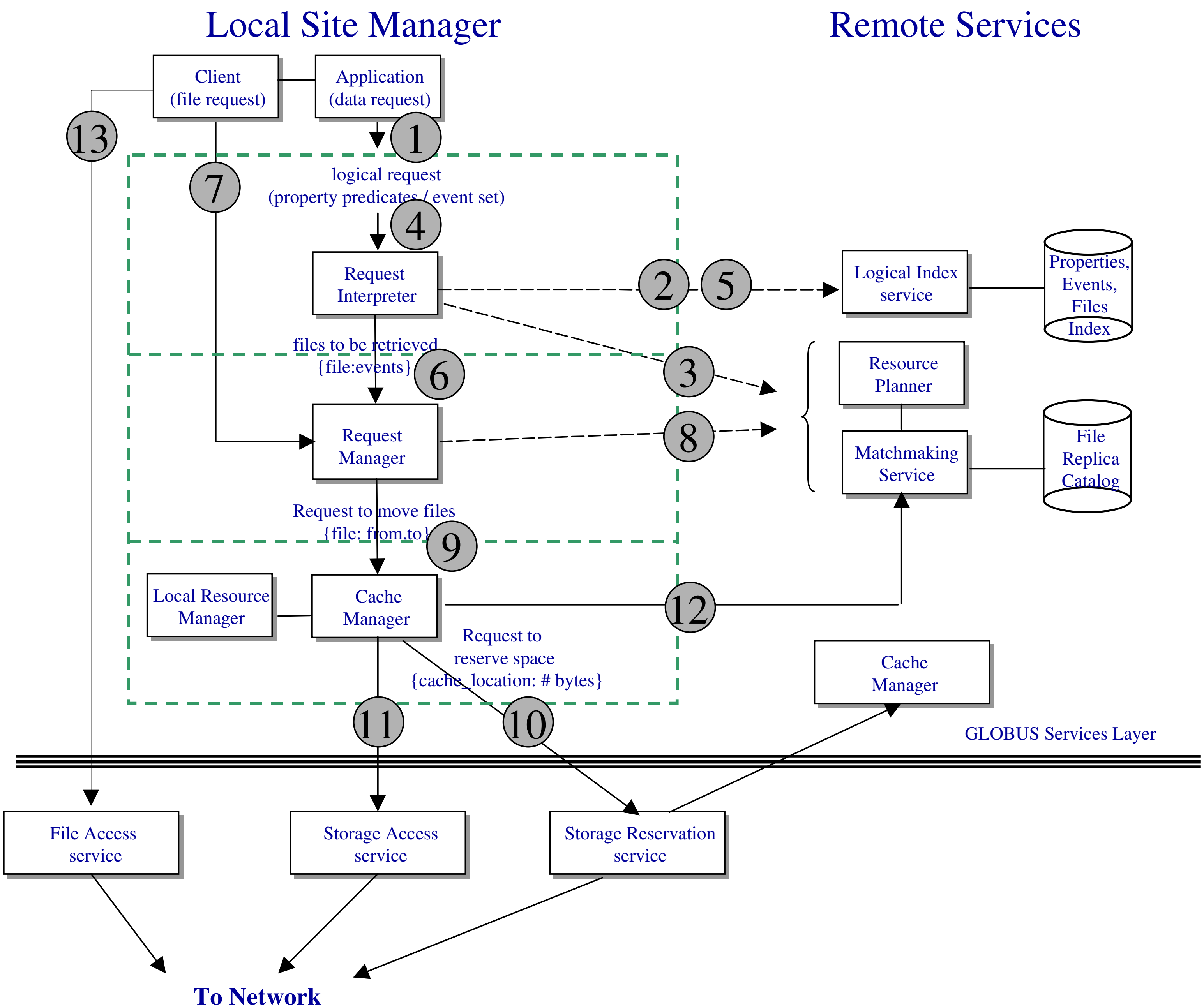


Fig 1: Architecture for the general scenario - needed APIs

PPDG First Year Milestones

• Project Start	August, 1999
• Decision on existing middleware to be integrated into the first-year Data Grid;	October, 1999
• First demonstration of high-speed site-to-site data replication;	January, 2000
• First demonstration of multi-site cached file access (3 sites);	February, 1999
• Deployment of high-speed site-to-site data replication in support of two particle-physics experiments;	July, 2000
• Deployment of multi-site cached file access in partial support of at least two particle-physics experiments.	August, 2000

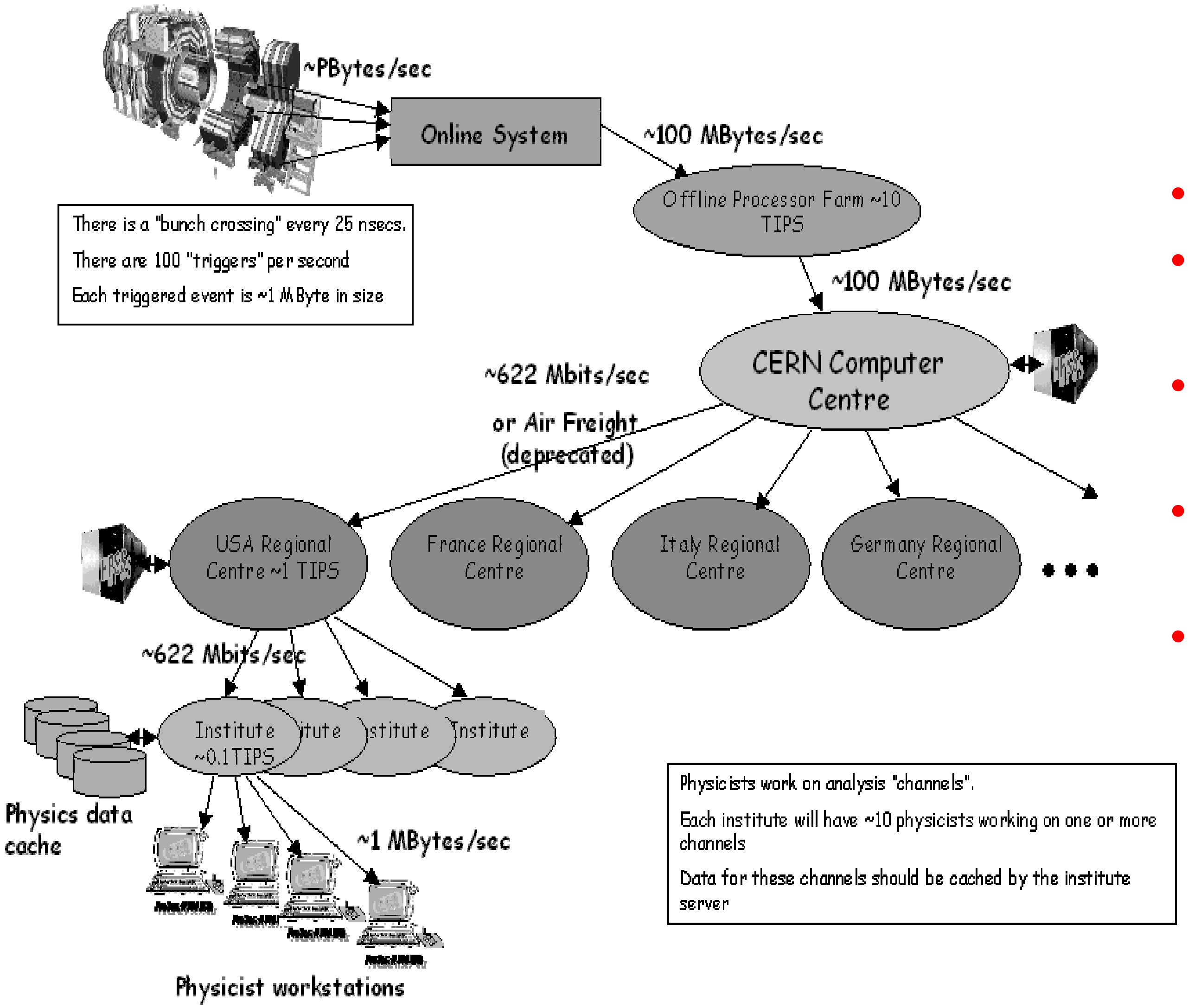
PPDG Longer-Term Goals

- Agent Computing

on

- Virtual Data

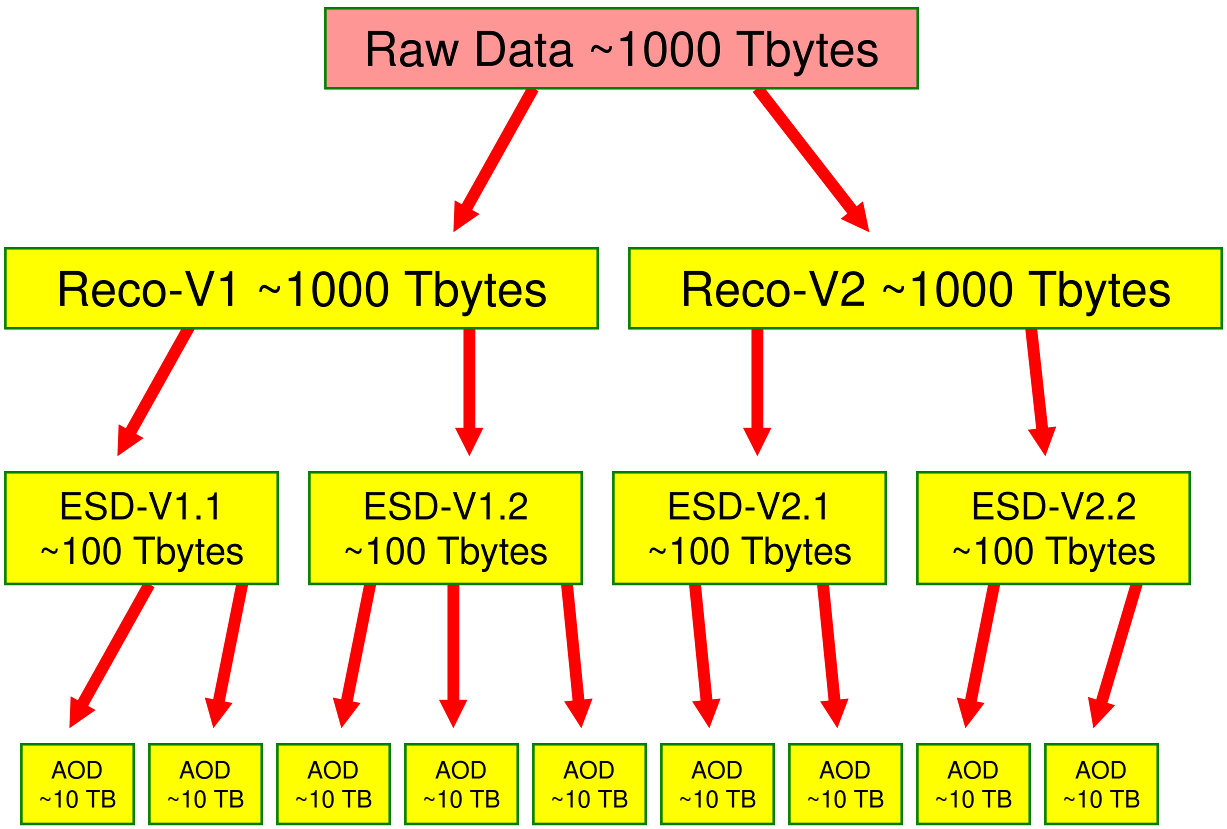
Why Agent Computing?



- **LHC Grid Hierarchy Example**
- **Tier0: CERN**
- **Tier1: National "Regional" Center**
- **Tier2: Regional Center**
- **Tier3: Institute Workgroup Server**
- **Tier4: Individual Desktop**
- **Total 5 Levels**

Why Virtual Data?

Typical particle physics experiment in 2000-2005:
On year of acquisition and analysis of data



Access Rates (aggregate, average)

100 Mbytes/s (2-5 physicists)

1000 Mbytes/s (10-20 physicists)

2000 Mbytes/s (~100 physicists)

4000 Mbytes/s (~300 physicists)

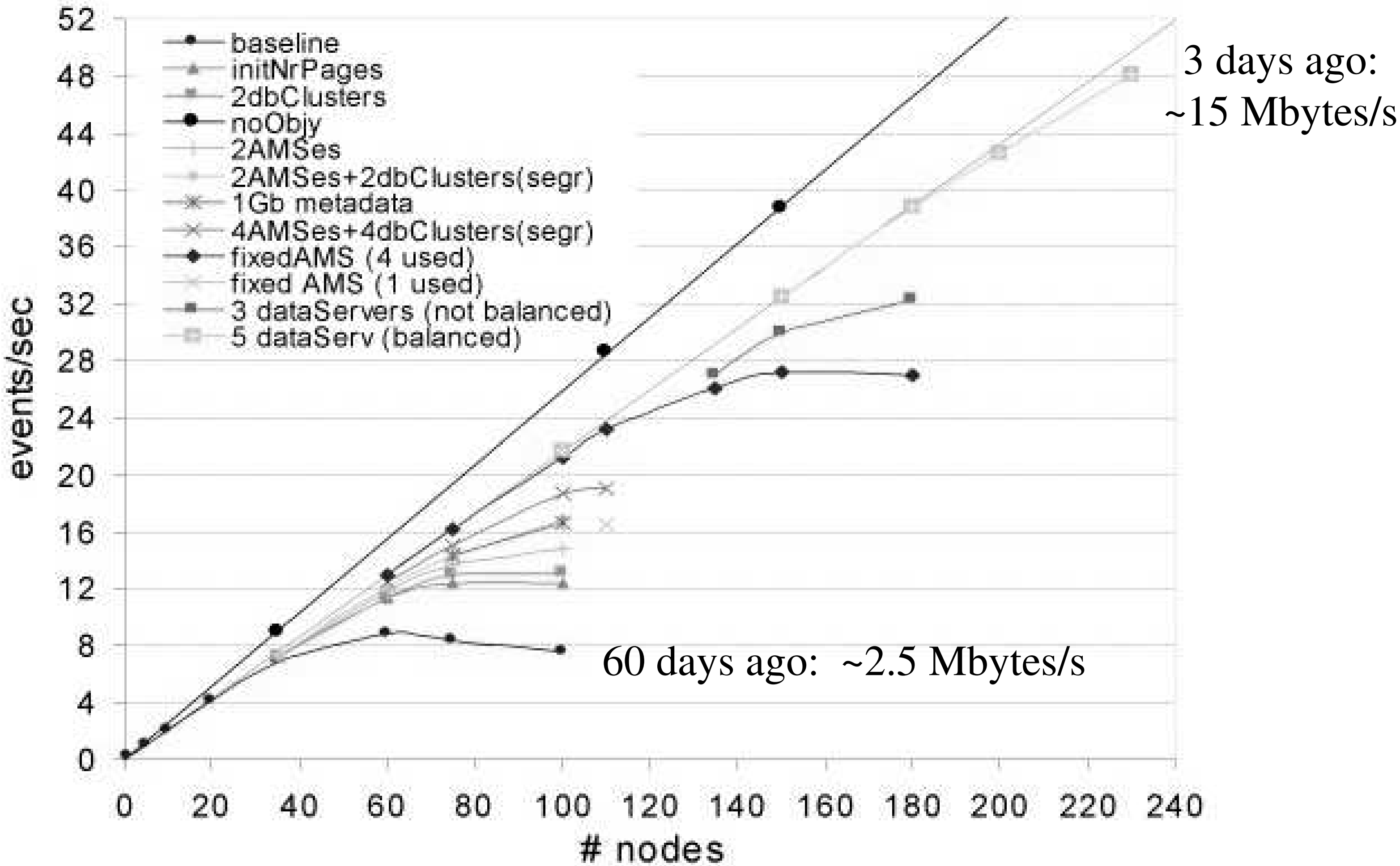
Existing Achievements

- SLAC-LBNL memory-to-memory transfer at 57 Mbytes/s over NTON (RIP[1]);
- Caltech tests of writing into Objectivity DB at 175 Mbytes/s

1) Not really moribund, but many decayed connections.

Cold Reality

(Writing into the BaBar Object Database at SLAC)



Testbed Requirements

- Site-to-Site Replication Service
  - 100 Mbyte/s goal possible through the resurrection of NTON (SLAC-LLNL-Caltech-LBNL are working on this).
- Multi-site Cached File Access System
  - Will use OC12, OC3, (even T3) as available
    - (even 20 Mits/s international links)
  - Need “Bulk Transfer” service:
    - Latency unimportant
    - Tbytes/day throughput important (Need prioritzed service to achieve this on international links)
    - Coexistence with other network users important. (This is the main PPDG need for differentiated services on ESnet)

How the PPDG Intends to Maximize its Impact and Relevance

- Develop and evolve a long-term plan for particle physics data management based on full exploitation of and contribution to relevant CS:
  - APOGEE draft proposal to DoE HENP (A Physics-Optimized Grid Environment for Experiments)
  - Proposed common project management for APOGEE and PPDG

APOGEE in Context