



HEP Computing

Matthias Kasemann
Fermilab

HEP computing: The next 5 years...(1)



- ◆ Data analysis for completed experiments continues
 - ◆ Challenges:
 - ◆ No major change to analysis model, code or infrastructure
 - ◆ Operation, continuity, maintaining expertise and effort

- ◆ Data collection and analysis for ongoing experiments
 - ◆ Challenges:
 - ◆ Data volume, compute resources, software organization
 - ◆ Operation, continuity, maintaining expertise and effort

HEP computing: The next 5 years...(2)



- ◆ Starting experiments:
 - ◆ Challenges:
 - ◆ Completion and verification of data and analysis model,
 - ◆ Data volume, compute resources, software organization, \$\$'s
 - ◆ Operation, continuity, maintaining expertise and effort

- ◆ Experiments in preparation:
 - ◆ Challenges:
 - ◆ Definition and implementation of data and analysis model,
 - ◆ data volume, compute resources, software organization, \$\$'s
 - ◆ continuity, getting and maintaining expertise and effort
 - ◆ Build for change: applications, data models...
 - ◆ Build compute models which are adaptable to different local environments

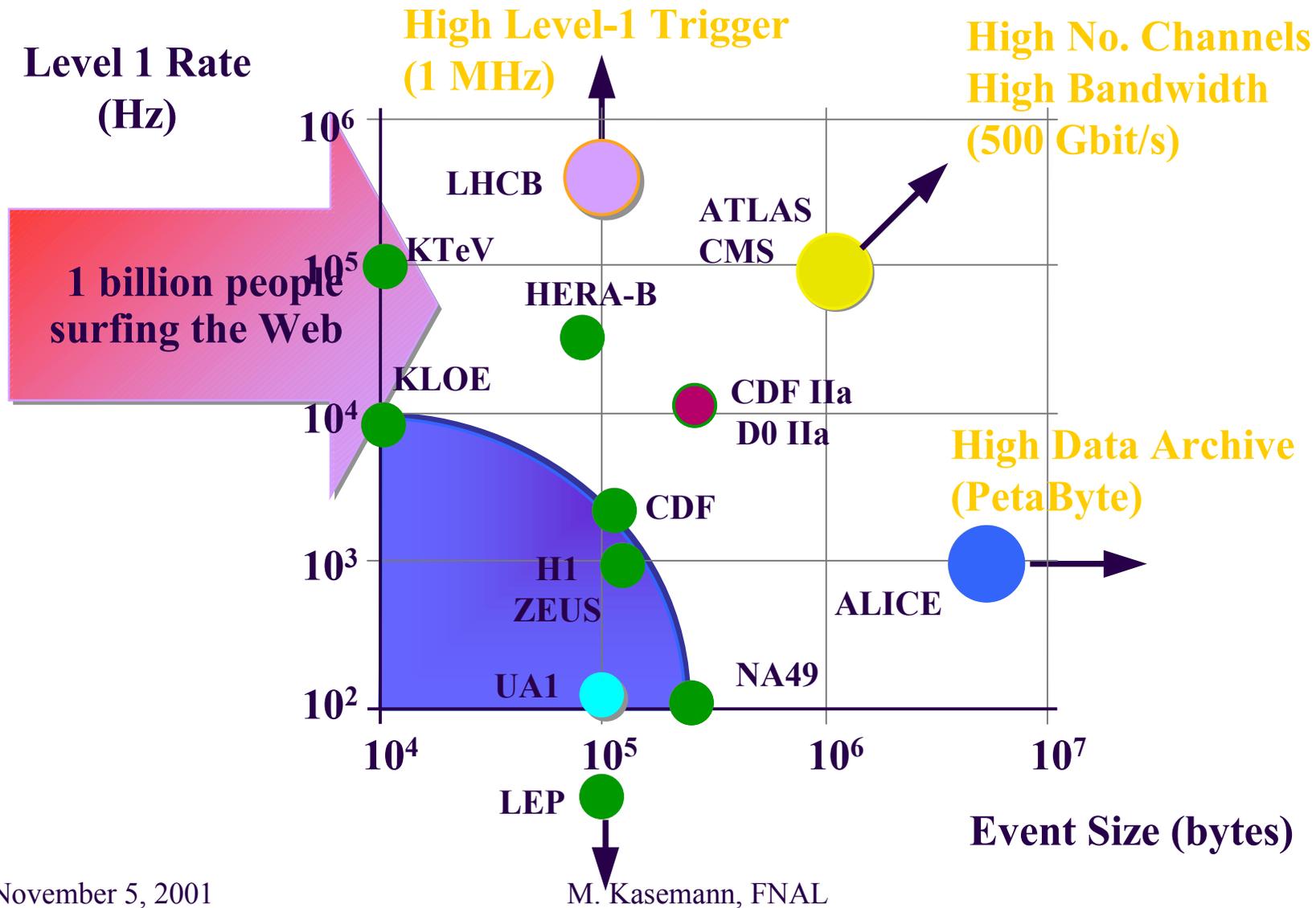
Run 2 Data Volumes



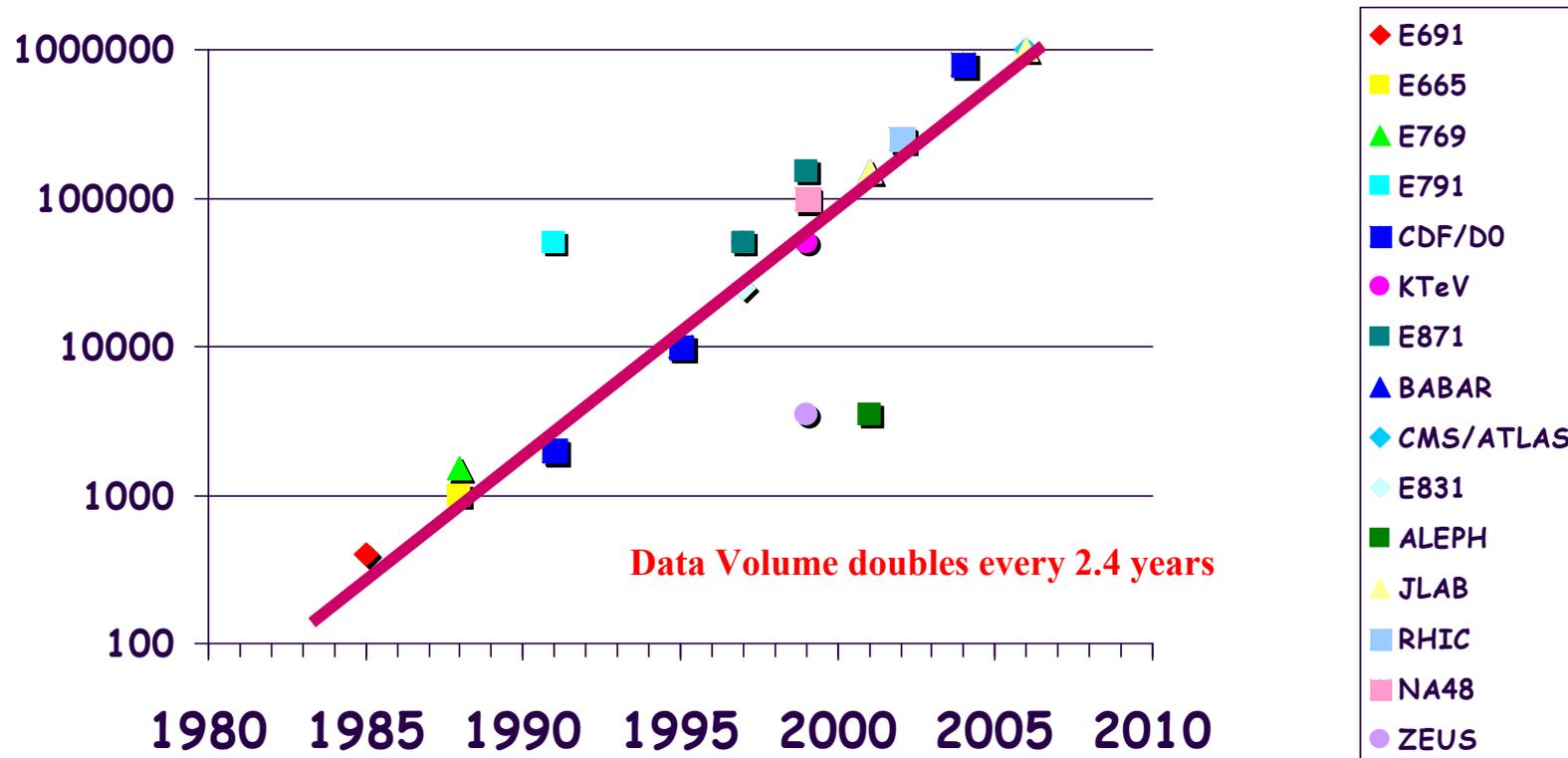
Category	Parameter	D0	CDF
DAQ rates	Peak rate	53 Hz	75 Hz
	Avg. evt. Size	250 KB	250 KB
	Level 2 output	1000 Hz	300 Hz
	maximum log rate	Scalable	80 BM/s
Data storage	# of events	600M/year	900 M/year
	RAW data	150 TB/year	250 TB/year
	Reconstructed data	75 TB/year	135 TB/year

- ◆ First Run 2b costs estimates based on scaling arguments
 - ◆ Use predicted luminosity profile
 - ◆ Assume technology advance (Moore's law)
 - ◆ CPU and data storage requirements both scale with data volume stored
- ◆ Data volume depends on physics selection in trigger
 - ◆ Can vary between 1 – 8 PB (Run 2a: 1 PB) per experiment
- ◆ Have to start preparation by 2002/2003

How Much Data is Involved?



Data Volume per experiment per year (in units of Gbytes)



Computing Needs: Comparison between Experiments



Experiment	Onsite CPU (SI95)	Onsite Disk (TB)	Onsite Tape (TB)	LAN Capacity	Data import/ export	Box count
500MHz PIII	20					
CMS	520,000	540	2000	46 GB/s	10 TB/day	~1400
CDF(Run2)	12,000	20	800	?		~250
D0(Run 2)	7,000	20	600	300 MB/s		~250
BaBar	10,000	10	300		0.5 TB/day	~300
CDF(Run1)	280	?	?	?		?
D0(Run 1)	295	1.5	65	300 Mb/s		180
ALEPH	300	?	5.5	1 Gb/s		1
DELPHI	515	1.2	?	1 Gb/s		20
L3	625	2	?	1 Gb/s	none	1
OPAL	835	1.6	?	1 Gb/s		1
NA45	587	1.3	2	1 Gb/s	5 GB/day	30
NA48	650	4.5	140	1 Gb/s	5 GB/day	50
KTeV	280	1	50	100 Mb/s	150 GB/day	2

Status: as of 1999

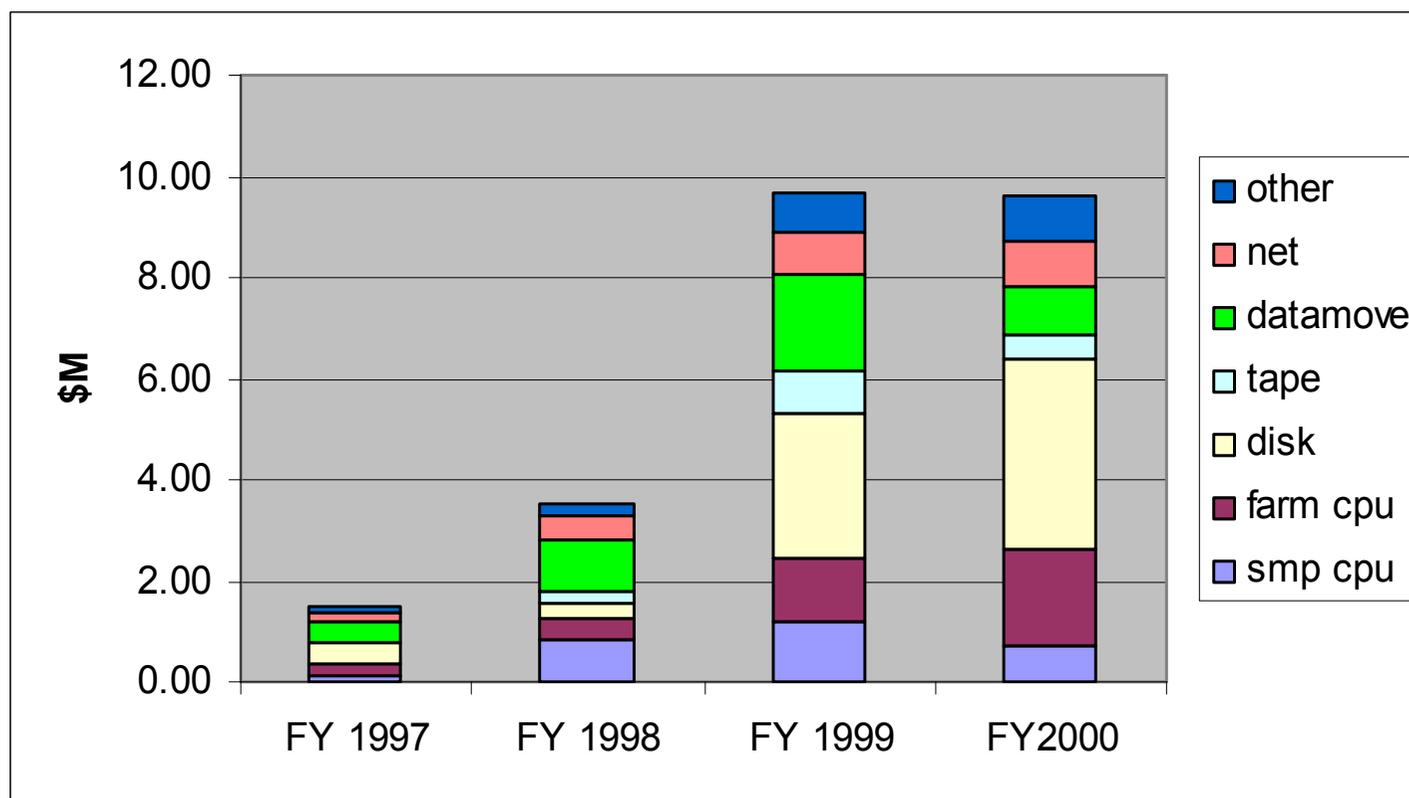
BaBar Offline Computing at SLAC:

Costs other than Personnel



(does not include "per physicist" costs such as desktop support, help desk, telephone, general site network)

Data Analysis for SLAC Physics
Richard P. Mount
CHEP 2000, Padova, February 2000



Does not
include
tapes

BaBar Offline Computing at SLAC:

Costs other than Personnel

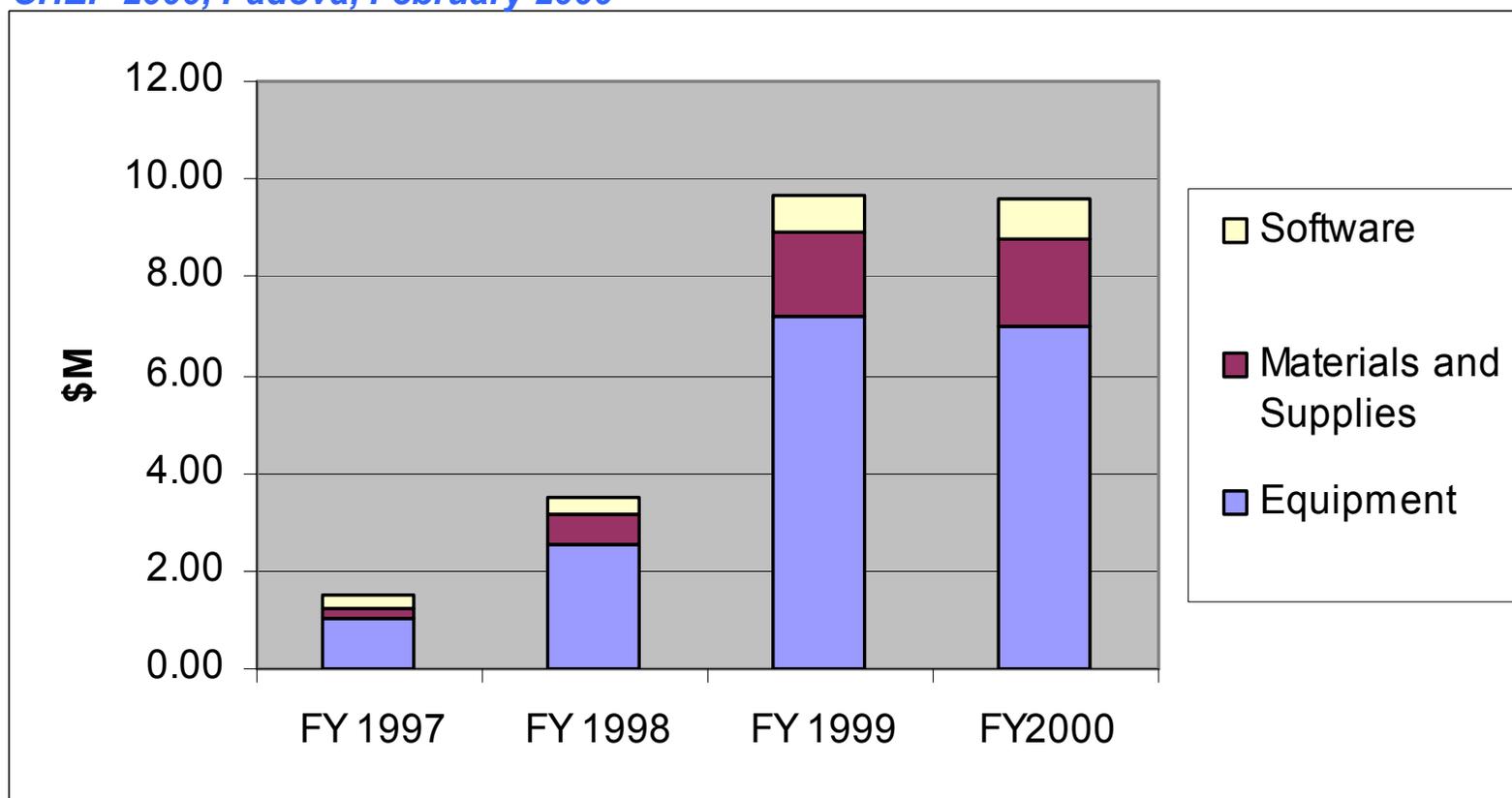


(does not include "per physicist" costs such as desktop support, help desk, telephone, general site network)

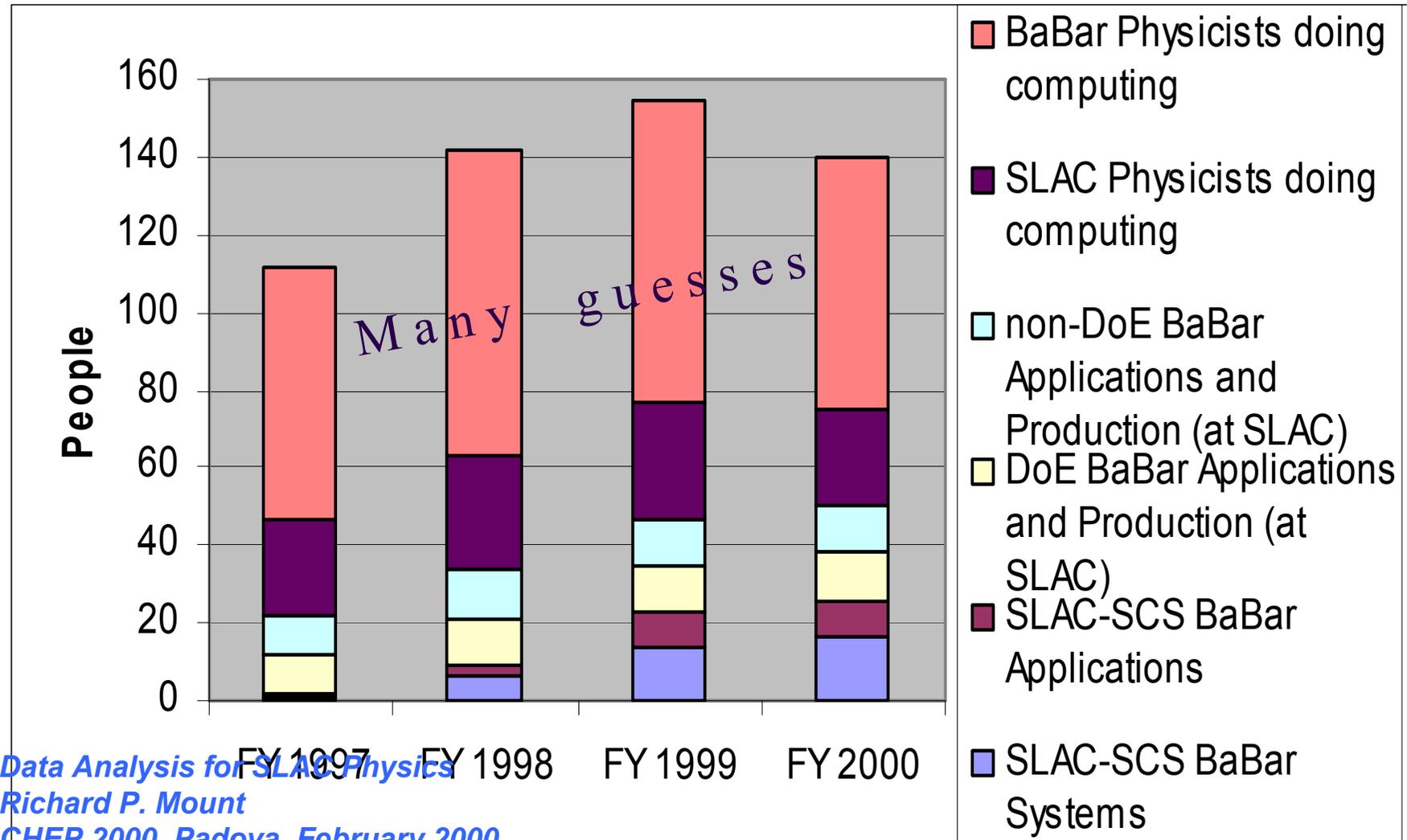
Data Analysis for SLAC Physics

Richard P. Mount

CHEP 2000, Padova, February 2000



BaBar Computing Personnel The Whole Story?



HEP computing: The next 5 years...(3)



- ◆ Challenges in big collaborations
 - ◆ Long and difficult planning process
 - ◆ More formal procedure required to commit resources
 - ◆ Long lifetime, need flexible solutions which allow for change
 - ◆ Any state of experiment longer than typical PhD or postdoc time
 - ◆ Need for professional IT participation and support

- ◆ Challenges in smaller collaborations
 - ◆ Limited in resources
 - ◆ Adapt and implement available solutions (“b-b-s”)

Hardware (aka "enough CPU")



- ◆ Explosion of number of farms installed
 - ◆ Very cost-effective
 - ◆ Linux is free; PC's are inexpensive
 - ◆ Interconnect: Fast/Giga Ethernet, Myrinet, Fibrechannel, even ATM
 - ◆ Despite recent growth, it's a mature process
 - ◆ Basic elements (PC, Linux, Network) are all mature technologies.
 - ◆ Problem solved.
 - ◆ But still left: Control & Monitor of thousands of (intelligent) things
 - ◆ But C&M does not seem to be a fundamental problem
- ◆ Conclusion on hardware: probably rightly skipped
 - ◆ It's the software that's harder to design, code and operate
 - ◆ And anyway the industry is many times better than us

Role of computer networking (1)



- ◆ State-of-the-art computer networking enables large international collaborations
 - ◆ needed for all aspects of collaborative work
 - ◆ to write the proposal,
 - ◆ produce and agree on the designs of the components and systems,
 - ◆ collaborate on overall planning and integration of the detector, confer on all aspects of the device, including the final physics results, and
 - ◆ provide information to collaborators and to the physics community and general public
 - ◆ Data from the experiment lives more-and-more on the network
 - ◆ All levels: raw, dst, aod, ntuple, draft-paper, paper

Role of computer networking (2)



- ◆ HEP developed its own national network in the early 1980s
- ◆ National research network backbones generally provide adequate support to HEP and other sciences.
- ◆ Specific network connections are used where HEP has found it necessary to support special capabilities that could not be supplied efficiently or capably enough through more general networks.
 - ◆ US-CERN, several HEP links in Europe...
- ◆ Dedicated HEP links are needed in special cases because
 - ◆ HEP requirements can be large and can overwhelm those of researchers in other fields
 - ◆ because regional networks do not give top priority to interregional connections

DØ: The Network is the Heart of the System



- ◆ Files are moved via LAN or WAN in the same manner – using various file transfer protocols over an IP packet network.
 - ◆ encp, bbftp (7 way parallel transfers), rcp, hpss form of cp, etc.
- ◆ Fermilab site sees greatest movement of data.
- ◆ Enstore file transfer protocol (encp) provides load balancing between multiple network interface cards
 - ◆ For Origin2000 each Gbit Ethernet interface needs 1 dedicated CPU and supports ~30MB/sec
- ◆ World-wide DØ Monte Carlo Production is up and working now
 - ◆ Current total Bandwidth to Fermilab ~50-100Mb/sec
 - ◆ Shipping MC data back and forth is essential
 - ◆ **Total Bandwidth ~200Mb/sec (2001)**
 - ◆ **Total Bandwidth ~400Mb/sec (2002)**
 - ◆ Real data processing at remote farms + reprocessing (?)
 - ◆ **Total Bandwidth ~800Mb/sec (2002)**
 - ◆ **Total Bandwidth ~(1200/1600/3200/4000)Mb/sec (2003/4/5/6)**
- ◆ Is Trans-Atlantic bandwidth available?

Run IIa Equipment Spending Profile

(Total for both CDF & D0 experiments)



- ◆ Mass storage: robotics, tape drives + interface computing.
- ◆ Production farms
- ◆ Analysis computers: support for many users for high statistics analysis (single system image, multi-CPU).
- ◆ Disk storage: permanent storage for frequently accessed data, staging pool for data stored on tape.
- ◆ Miscellaneous: networking, infrastructure, ...

Fiscal Year	MSS	Farms	Analysis	Disk	Misc	Total (both)
Spent in FY98	\$1.2M	\$200K	-	\$200K	\$400K	\$2M
Spent in FY99	\$2.2M	\$700K	\$2M	\$800K	\$300K	\$6M
Spent in FY00	\$450K	\$350K	\$100K	\$300K	\$800K	\$2M
Budget FY01	\$450K	\$350K	\$2.14M	\$690K	\$70K	\$4M
Plan for FY02	\$500K	\$1.2M	\$2.16M	\$610K	\$30K	\$4.2M
Total Needs	\$4.8M	\$2.8M	\$6.4M	\$2.6M	\$1.6M	\$18.2M
Continuing Operations (FY2002 and beyond)						\$2M

Data analysis in international collaborations: past



- ◆ In the past analysis was centered at the experimental site
 - ◆ a few major external centers were used.
 - ◆ Up the mid 90s bulk data were transferred by shipping tapes, networks were used for programs and conditions data.
 - ◆ External analysis centers served the local/national users only.
 - ◆ Often staff (and equipment) from the external center being placed at the experimental site to ensure the flow of tapes.
 - ◆ The external analysis often was significantly disconnected from the collaboration mainstream.

Data analysis in international collaborations: truly distributed



- ◆ Why?
 - ◆ For one experiment looking ahead for a few years only centralized resources may be most cost effective, but:
 - ◆ national and local interests leads to massive national and local investments
 - ◆ For BaBar:
 - ◆ The total annual value of foreign centers to the US-based program is greatly in excess of the estimated cost to the US of creating the required high-speed paths from SLAC to the landing points of lines WAN funded by foreign collaborators
 - ◆ Future world-scale experimental programs must be planned with explicit support for a collaborative environment that allows many nations to be full participants in the challenges of data analysis.

Distributed computing:



- ◆ Networking is an expensive resource, use must be well planned
- ◆ Pre-emptive transfers can be used to improve responsiveness at the cost of some extra network traffic.
- ◆ Multi-tiered architecture must become more general and flexible
 - ◆ to accommodate the very large uncertainties in the relative costs of CPU, storage and networking
 - ◆ To enable physicists to work effectively in the face of data having unprecedented volume and complexity
- ◆ Aim for transparency and location independence of data access
 - ◆ the need for individual physicists to understand and manipulate all the underlying transport and task-management systems would be too complex

Distributed Computing

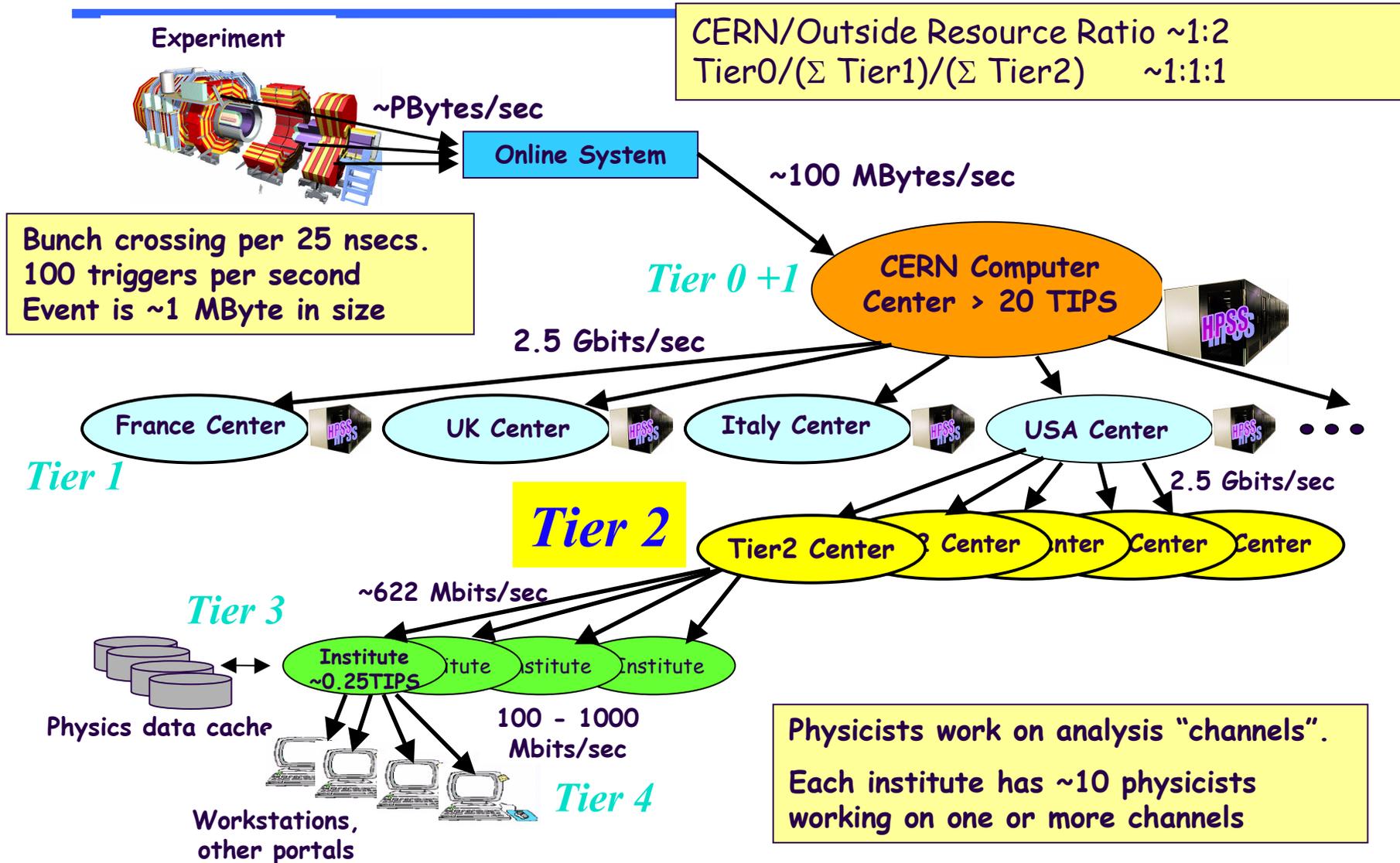


6/13/01: **The New York Times**
ON THE WEB

"It turns out that distributed computing is really hard," said Eric Schmidt, the **chairman of Google**, the Internet search engine company.

"It's much harder than it looks. It has to work across different networks with different kinds of security, or otherwise it ends up being a single-vendor solution, which is not what the industry wants."

Example: CMS Data Grid



Tier1 and Tier2 Centers



- ◆ Tier1 centers
 - ◆ National laboratory scale: large CPU, disk, tape resources
 - ◆ High speed networks
 - ◆ Many personnel with broad expertise
 - ◆ Central resource for large region
- ◆ Tier2 centers
 - ◆ New concept in LHC distributed computing hierarchy
 - ◆ Size \approx [national lab * university]^{1/2}
 - ◆ Based at large University or small laboratory
 - ◆ Emphasis on small staff, simple configuration & operation
- ◆ Tier2 role
 - ◆ Simulations, analysis, data caching
 - ◆ Serve small country, or region within large country

Why Regional Centers?



- ◆ Bring computing facilities closer to home
 - ◆ final analysis on a compact cluster in the physics department
- ◆ Exploit established computing expertise & infrastructure
- ◆ Reduce dependence on links to CERN
 - ◆ full ESD available nearby - through a fat, fast, reliable network link
- ◆ Tap funding sources not otherwise available to HEP
- ◆ Devolve control over resource allocation
 - ◆ national interests?
 - ◆ regional interests?
 - ◆ at the expense of physics interests?

Regional Centers Services and Facilities



- ◆ Regional Centers will
 - ◆ Provide all technical services and data services required to do the analysis
 - ◆ Maintain all (or a large fraction of) the processed analysis data. Possibly may only have large subsets based on physics channels. Maintain a fixed fraction of fully reconstructed and raw data
 - ◆ Cache or mirror the calibration constants
 - ◆ Maintain excellent network connectivity to CERN and excellent connectivity to users in the region. Data transfer over the network is preferred for all transactions but transfer of very large datasets on removable data volumes is not ruled out.
 - ◆ Share/develop common maintenance, validation, and production software with CERN and the collaboration

Regional Centers Services and Facilities



- ◆ Provide services to physicists in the region, contribute a fair share to post-reconstruction processing and data analysis, collaborate with other RCs and CERN on common projects, and provide services to members of other regions on a best effort basis to further the science of the experiment
- ◆ Provide support services, training, documentation, trouble shooting to RC and remote users in the region

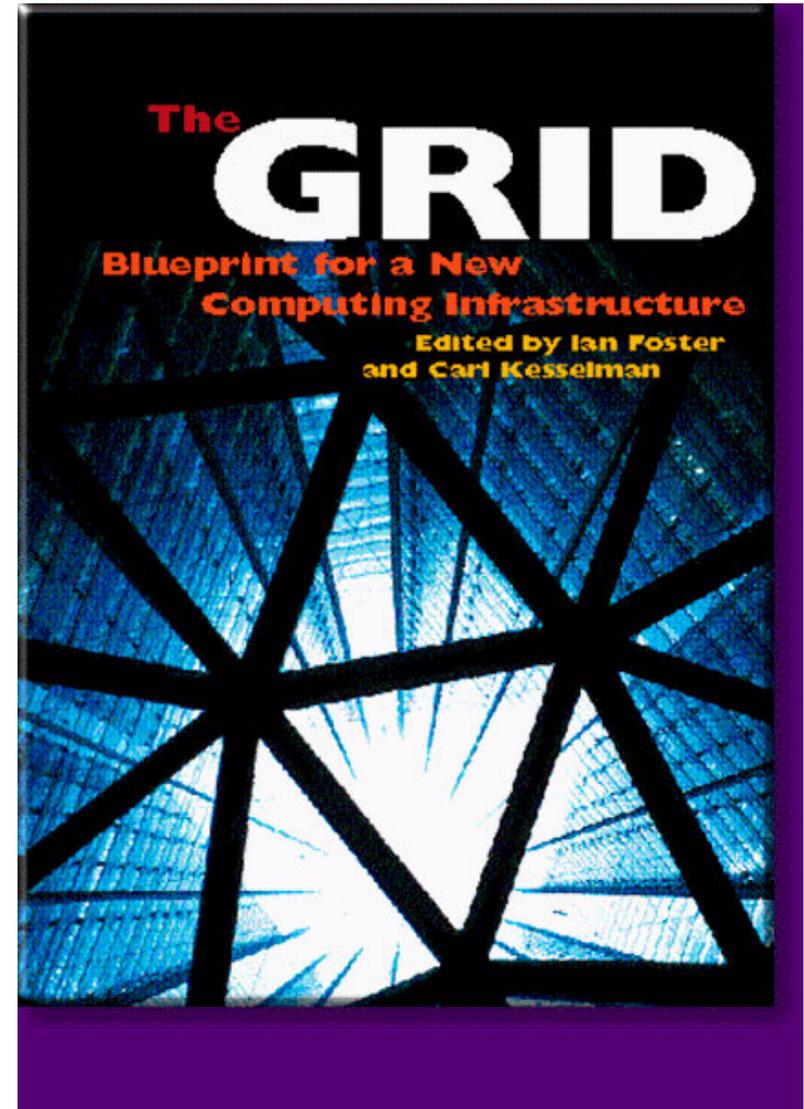
Motivations for Regional Centers



- ◆ To maximize the intellectual contribution of physicists all over the world without requiring their physical presence at CERN
- ◆ Acknowledgement of possible limitations of network bandwidth
- ◆ A way of utilizing the expertise and resources residing in computing centers all over the world
- ◆ Allows people to make choices on how they analyze data based on availability or proximity of various resources such as CPU, data, or network bandwidth.

Information Grids: the solution to the LHC Data Challenge ?

- ◆ Next step after Web/Internet
- ◆ Information Sockets dynamically deliver data and computational resources
- ◆ Analogy to the Electric Grid
- ◆ Major difference:
All electrons are similar...
All bits of information are not.
- ◆ Hot research topic



One View of Requirements



Carl Kesselman

Center for Grid Technologies

USC/Information Sciences Institute

- ◆ Identity & authentication
- ◆ Authorization & policy
- ◆ Resource discovery
- ◆ Resource characterization
- ◆ Resource allocation
- ◆ (Co-)reservation, workflow
- ◆ Distributed algorithms
- ◆ Remote data access
- ◆ High-speed data transfer
- ◆ Performance guarantees
- ◆ Monitoring
- ◆ Adaptation
- ◆ Intrusion detection
- ◆ Resource management
- ◆ Accounting & payment
- ◆ Fault management
- ◆ System evolution
- ◆ Etc.
- ◆ Etc.
- ◆ ...

Another View: "Three Obstacles to Making Grid Computing Routine"



Carl Kesselman

Center for Grid Technologies
USC/Information Sciences Institute

- New approaches to problem solving
 - ◆ Data Grids, distributed computing, peer-to-peer, collaboration grids, ...

- Structuring and writing programs

- ◆ Abstractions, tools

Programming Problem

- Enabling resource sharing across distinct institutions

- ◆ Resource discovery, access, reservation, allocation; authentication, authorization, policy; communication; fault detection and notification; ...

Systems Problem

Aspects of the Systems Problem



Carl Kesselman

Center for Grid Technologies
USC/Information Sciences Institute

- Need for interoperability when different groups want to share resources
 - ◆ Diverse components, policies, mechanisms
 - ◆ E.g., standard notions of identity, means of communication, resource descriptions
- Need for shared infrastructure services to avoid repeated development, installation
 - ◆ E.g., one port/service/protocol for remote access to computing, not one per tool/appln
 - ◆ E.g., Certificate Authorities: expensive to run
- ◆ A common need for protocols & services

Protocol-Oriented View of Grid Architecture



Carl Kesselman

Center for Grid Technologies
USC/Information Sciences Institute

- ◆ Development of Grid protocols & services
 - ◆ Protocol-mediated access to remote resources
 - ◆ New services: e.g., resource brokering
 - ◆ “On the Grid” = speak Intergrid protocols
 - ◆ Mostly (extensions to) existing protocols
- ◆ Development of Grid APIs & SDKs
 - ◆ Facilitate application development by supplying higher-level abstractions
- ◆ The (hugely successful) model is the Internet

Layered Grid Architecture (By Analogy to Internet Architecture)



Carl Kesselman

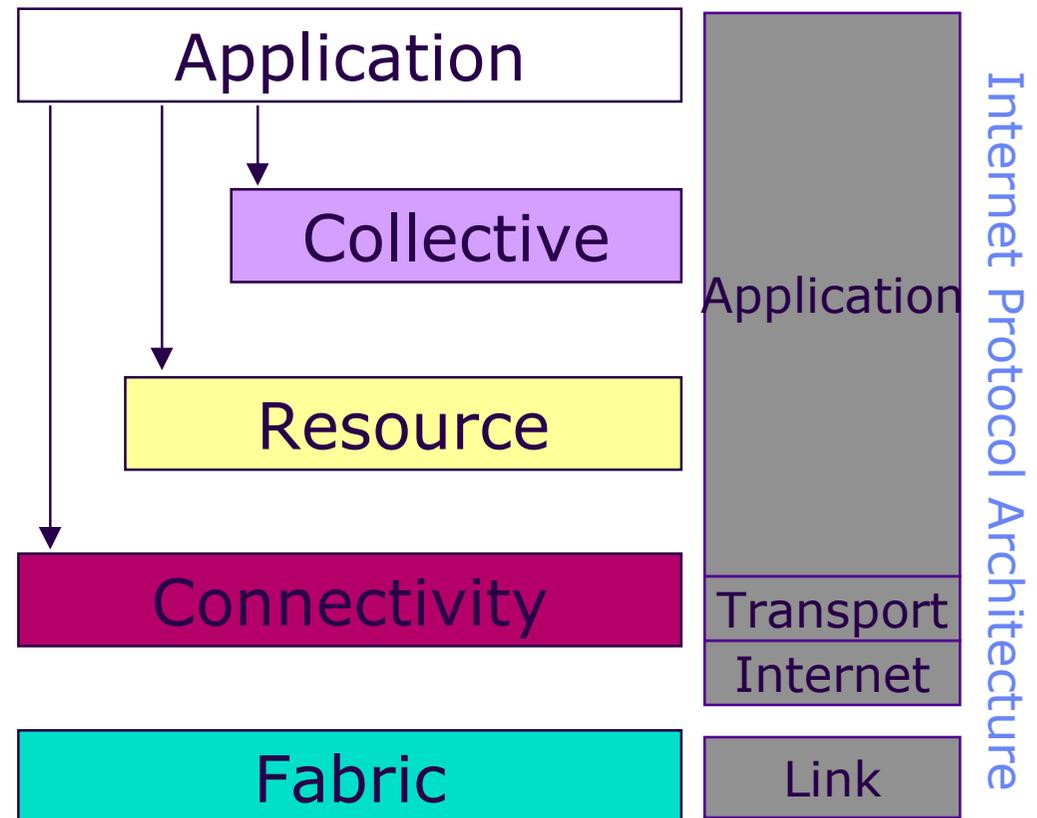
Center for Grid Technologies
USC/Information Sciences Institute

“Coordinating multiple resources”:
ubiquitous infrastructure services,
app-specific distributed services

“Sharing single resources”:
negotiating access, controlling use

“Talking to things”:
communication (Internet protocols) & security

“Controlling things locally”:
Access to, & control of, resources





Context: Major Data Grid Projects

➤ Funded projects

➔ GriPhyN	USA	NSF, \$11.9M + \$1.6M
➔ PPDG I	USA	DOE, \$2M
➔ PPDG II	USA	DOE, \$9.5M
➔ EU DataGrid	EU	\$9.3M

➤ Proposed projects

➔ iVDGL	USA	NSF, \$15M + \$1.8M + UK
➔ DTF	USA	NSF, \$45M + \$4M/yr
➔ DataTag	EU	EC, \$3M?
➔ GridPP	UK	PPARC, > \$15M
➔ AstroGrid	UK	PPARC, > \$2M

➤ Other national projects

- ➔ UK e-Science (> \$100M for 2001-2004)
- ➔ Italy, France, Japan ?
- ➔ EU networking initiatives (Géant, Danté, SURFNet)



HENP Intergrid Coordination

- Mechanism for communication and coordination across the HENP Data Grid projects. Common efforts in three major areas:
 - ➔ InterGrid Coordination Board (HICB) for high level coordination
 - ➔ Joint Technical Board (JTB)
 - ➔ Common Projects and Task Forces to address needs in specific technical areas
- Why only HENP?
 - ➔ Common interests for delivery of working applications of the experiments and already established milestones.
 - ➔ Cross-cut view of HENP requirements is essential to ensure compatibility of deliverables.

Some more thoughts



- ◆ Computing for HEP experiments is costly
 - ◆ In \$\$'s, people and time
 - ◆ Need R&D, prototyping and test-beds to develop solutions and validate choices

need to do a better job here
 - ◆ Improving the engineering aspect of computing for HEP experiments is essential
 - ◆ Treat computing and software as a project (see www.pmi.org):
 - ◆ Project lifecycles, milestones, resource estimates, reviews
 - ◆ Documenting conditions and work performed is essential for success
 - ◆ Track detector building for 20 years
 - ◆ Log data taking and processing conditions
 - ◆ Analysis steps, algorithms, cuts
- } As transparent and automatic as possible

Core Software developers needed for LHC



Year	2000 have (missing)	2001	2002	2003	2004	2005
ALICE	12(5)	17.5	16.5	17	17.5	16.5
ATLAS ¹	23(8)	36	35	30	28	29
CMS	15(10)	27	31	33	33	33
LHCb	14(5)	25	24	23	22	21
Totals	64(28)	105.5	106.5	103	100.5	99.5

LHC computing: challenges



- ◆ perform data challenges of increasing size and complexity
- ◆ Current cost estimates based on forecast evolution of price and performance of computer hardware
- ◆ hardware costs of initial set-up of LHC distributed computer centres (Tier-0 to -2):
 - ◆ 240 MCHF
 - ◆ CERN-based Tier-0+1 centre: about 1/3 of total.
- ◆ investment for initial system to be spent in 2005, 2006 and 2007, in ~ equal portions
 - ◆ (assuming LHC start-up in 2006 and reach of design luminosity in 2007)
- ◆ Materials & Operation of LHC computing system:
 - ◆ rolling replacement within constant budget
 - ◆ requires ~ 1/3 of initial investment per year (~ 80 MCHF world-wide) - includes steady evolution of capacity
- ◆ set-up of a common prototype as joint project (experiments, CERN-IT, major regional centres),
 - ◆ reaching ~50% of overall computing structure of 1 LHC experiment by ~2003/4

... and some more



- ◆ I am convinced:
 - ◆ that computing for HEP is as important as building the detector
 - ◆ that computing must be planned and ‘projectized’
 - ◆ that computing cannot be done by physicist alone
 - ◆ that computing cannot be done by “professionals” alone
 - ◆ work on software and computing must start early

- ◆ There must/should be a place in physics for building HEP computing and analysis systems as well as building detectors
 - ◆ Corollary:

- ◆ There must be a career for physicists working on computing as there is a career for physicists building detectors

Relying on experts



- ◆ In some cases we are trying to play “computer scientist”
 - ◆ We shouldn't. We should leave this task to computer scientists, i.e. professionals. At least for the core software.
- ◆ We have done that already with the big detectors
 - ◆ I would not work on an experiment where the mechanics of the magnet is designed by a “jack of all trades” HEPhysicist who learned it on the job.
 - ◆ Unless the HEPhysicist was a uniquely gifted person
 - ◆ Complexity (detector and computing) has overtaken the average HEPhysicist
 - ◆ Engineers are now necessary; we can work with them; guide them; help them; disagree with them

Computing for HEP: The full costs?



- ◆ Space
- ◆ Power, cooling
- ◆ Software
- ◆ Computing Hardware

- ◆ LAN
- ◆ Replacement/Expansion 30% per year

- ◆ Mass storage

- ◆ People