

# **Research Computing and Grid Technology: How These Developments May Impact Higher Education I/T Services**

**Patrick Dreher**

**Research Scientist & Associate Director  
MIT Laboratory for Nuclear Science**

**Common Solutions Group Meeting  
Seattle, Washington  
September 19, 2002**

# Outline

- **Traditional model for university research**
- **Forces driving university research beyond the traditional model**
- **Examples of leading-edge research computing projects and the rise of grid technology**
- **Vision for the immediate future:**
  - **Horizons for research computing for the first decade of the 21<sup>st</sup> Century**
  - **Implications for university I/T**

# **Traditional Model for Faculty Research at a University**

- **Many research projects at universities follow a model of an individual professor or group of researchers working on a project at a university lab**
- **Computational requirements for the research group are handled through a combination university-wide I/T possibly supplemented by individual research group resources**
- **The research group usually depends on the campus network for individual file transfers and e-mail connectivity with collaborators**

# **Beyond the Traditional Model: How Did We Get To The Present State of Research Computing?**

- **Increase in faculty hiring at research universities in the 2<sup>nd</sup> half of the 20<sup>th</sup> Century yielded a geographically distributed set of researchers**
- **These researchers began collaborations that differed in scope from their individual investigator counterparts**
- **Rise of “big science” projects in the late 20<sup>th</sup> Century now involve dozens to thousands of collaborators internationally**

# **Beyond the Traditional Model: How Did We Get To The Present State of Research Computing? (cont'd)**

- **Developments in technology now allow “big science” projects to construct the necessary instruments for this research**
- **These equipment needs are now larger than any one university can underwrite**
- **Now have geographically dispersed researchers using instrument(s) at remote sites for a common project**

# **Paradigm for Scientific Research in the 21<sup>st</sup> Century**

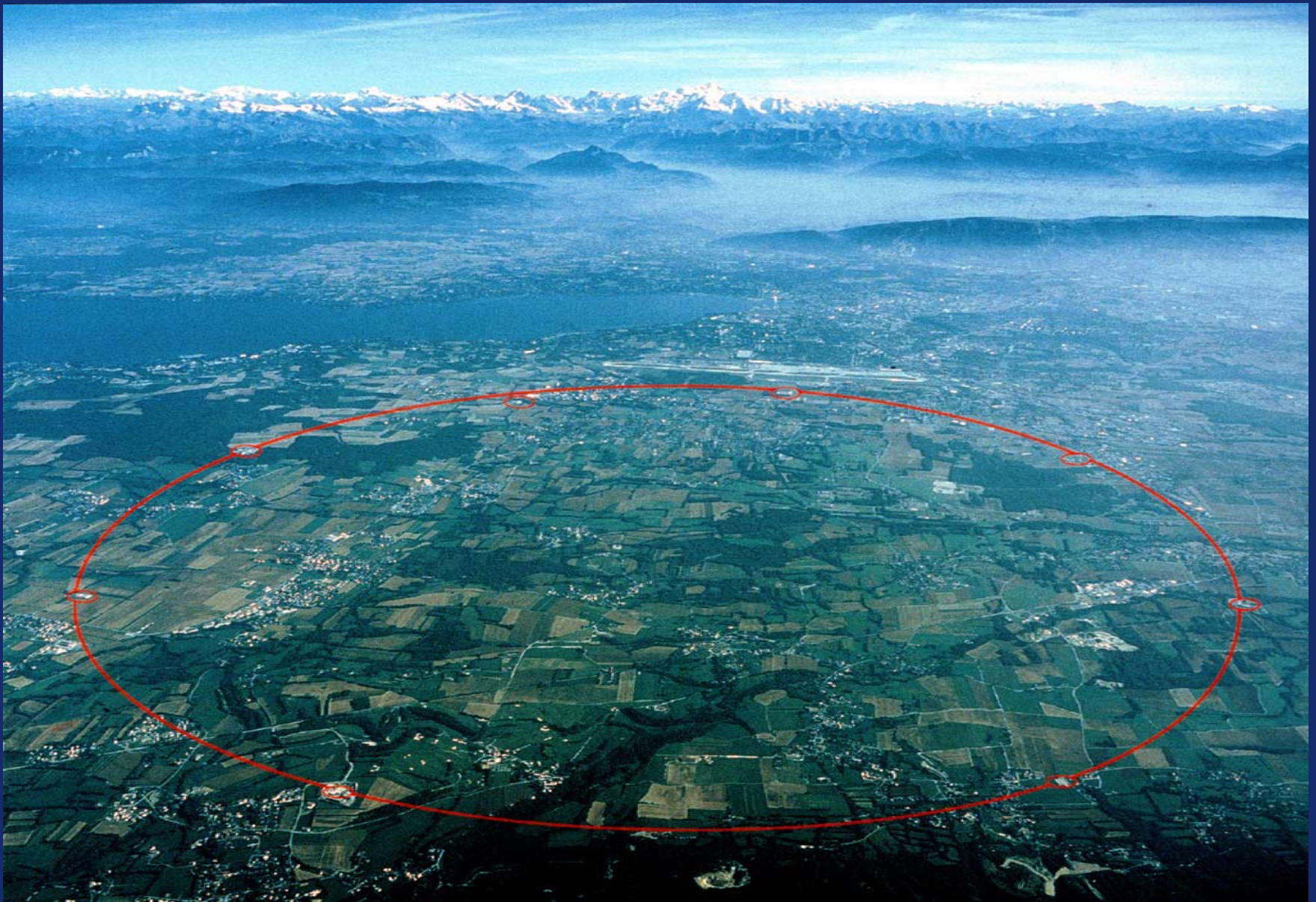
- **No one location will contain all of the resources needed for these projects**
- **Trend is toward more science and engineering research collaborations that will increasingly have the characteristics**
  - **multi-discipline**
  - **multi-institution**
  - **multi-instrument**
  - **intellectual and computing capacity distributed among participating sites that will span international boundaries**

# Examples of Funded Research

## Computing Projects Under Construction

- **Examples of scientific projects with different research computing needs and requirements being designed and constructed at the present time**
  - High Energy Physics -- Large Hadron Collider
  - National Virtual Observatory
  - Theoretical physics calculations
- **Time scale of 3 – 5 years before these projects are in production generating scientific data**
- **Development of grid technology as an integral component of these projects**
- **How will projects of this type impact and change university I/T planning and deliverables**

# **Experimental High Energy Physics: The Large Hadron Collider Project**



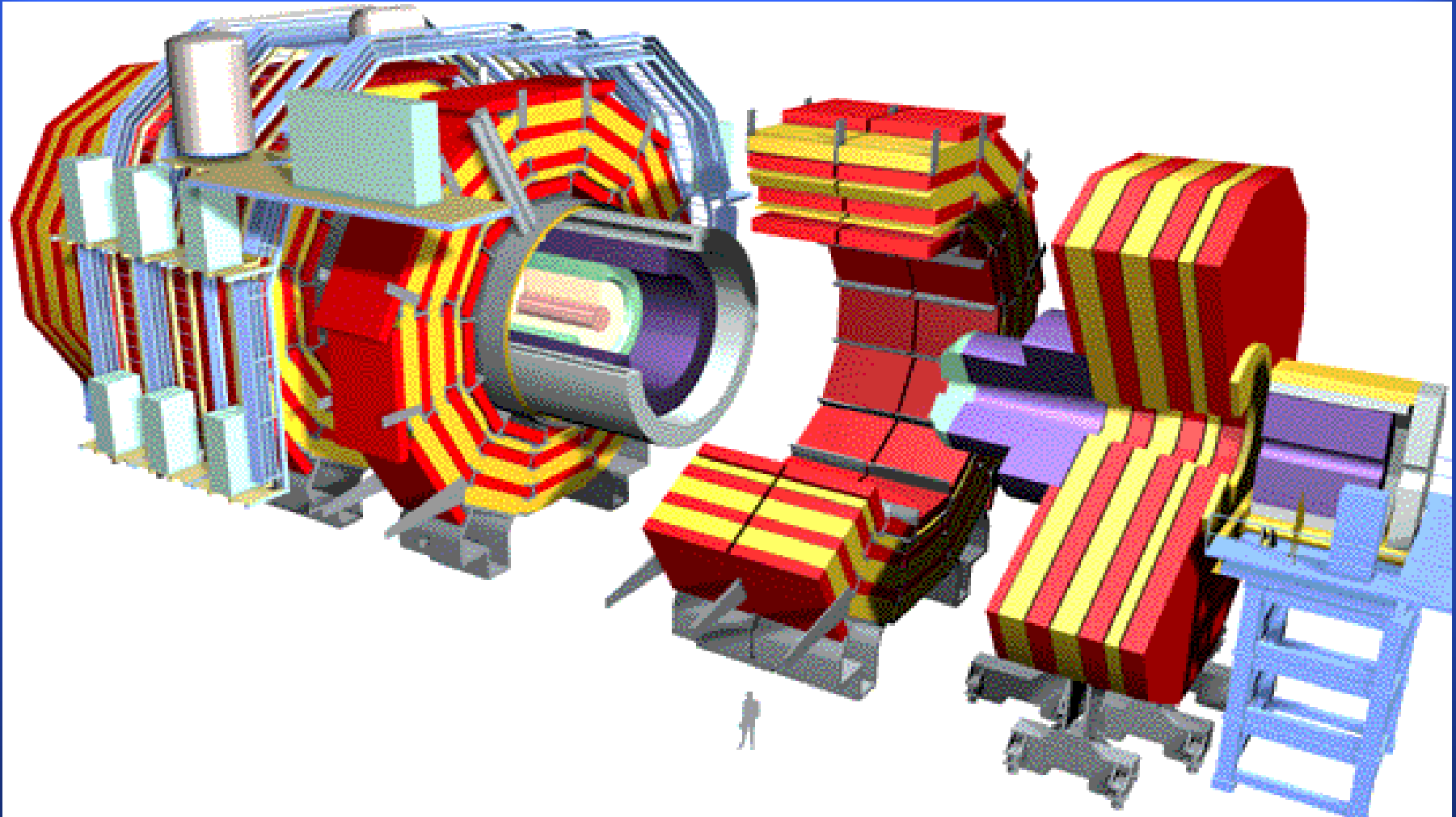
CERN photo

Common Solutions Group  
September 19, 2002

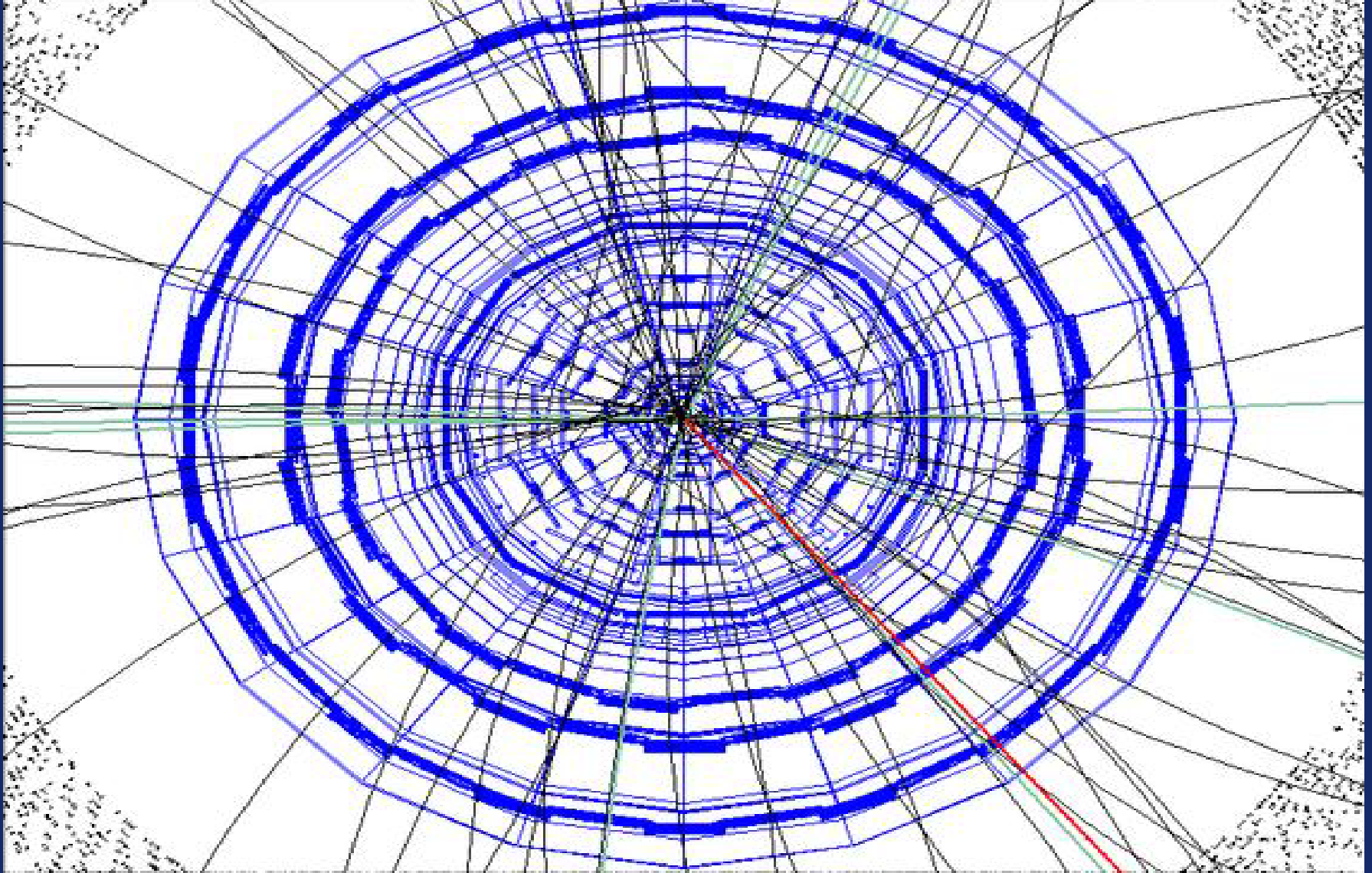
Research Computing and Grid Technology:  
How These Developments May Impact Higher Education  
I/T Services  
Patrick Dreher, MIT

09/02 ID100-9

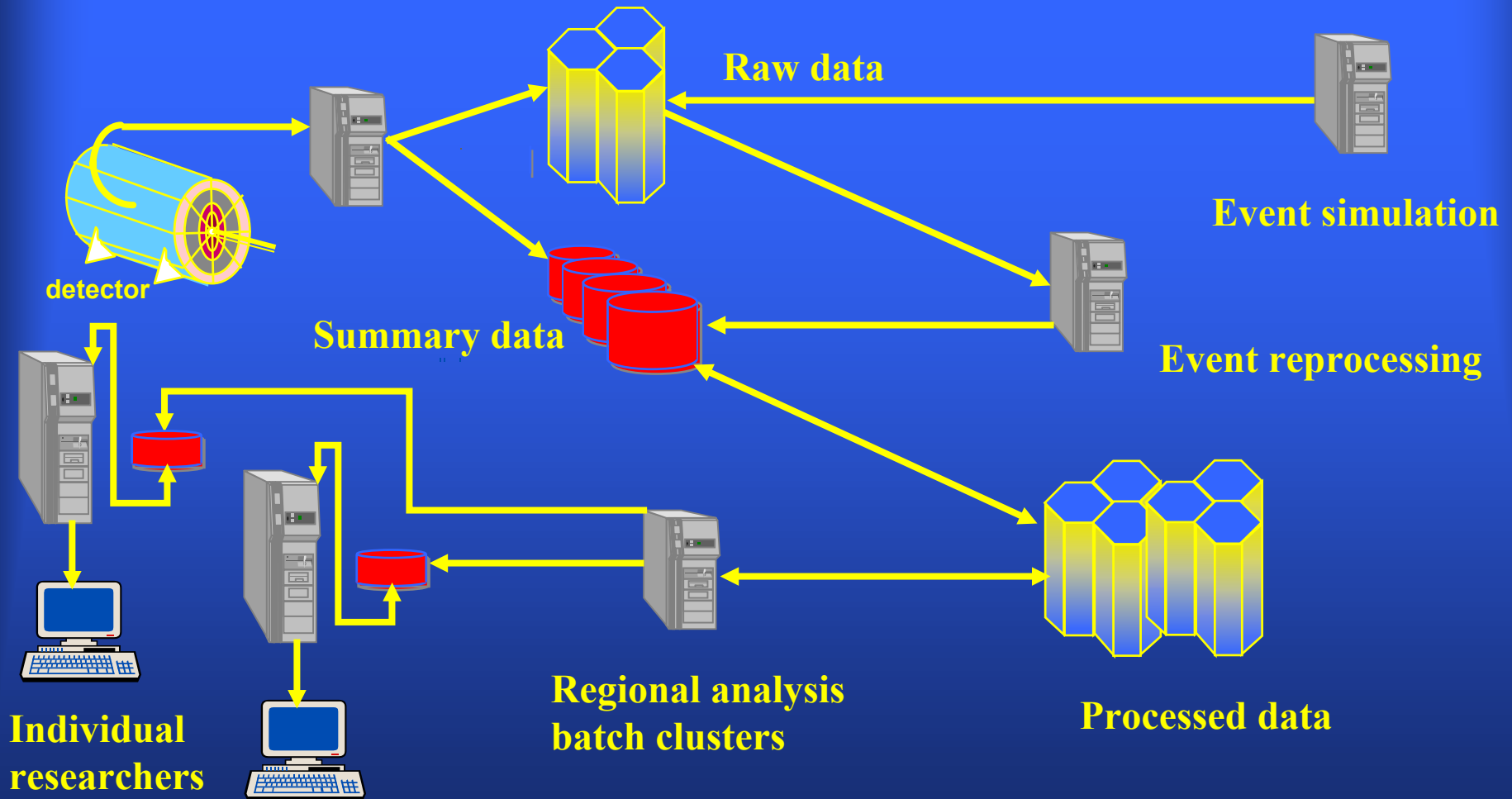
# One of the LHC Detectors



Event : 6512 Run : 151240 EventType : DATA | Unpresc: 41,11,43,44,13,17,19,21,23,30 Presc: 44,17 Myron mode: 0



# Computational Components



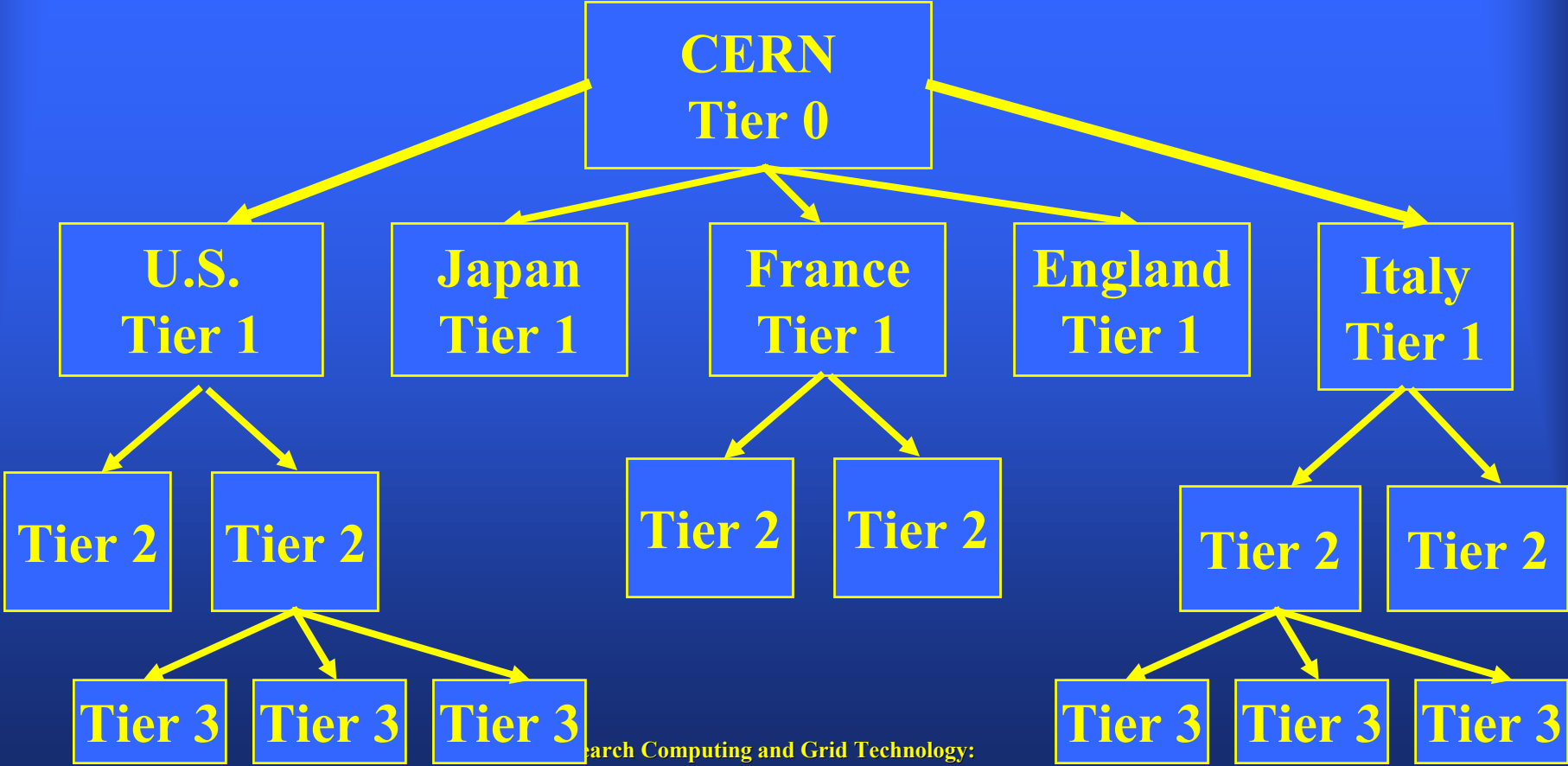
# Data Acquisition Requirements

- Large aggregate requirements for data acquisition, storage, distribution and computation
- Data acquisition 1 – 100 GB/sec
- Storage
  - Raw recording rate .1 – 1 GB/sec
  - Incremental tape storage increase of 5 – 8 Petabytes/year
- Large numbers of independent events requiring synchronization with metadata for detector parameters

# Operational Characteristics

- **Several thousand collaborators world-wide for each experimental detector requiring synchronized world-wide data distribution**
- **Grid technology must be implemented to coordinate, process, and analyze the experimental data**

# Distributed Computing for HEP



# Computing Capacity Required for the Experiments in 2007\*

	Tier 0	Tier 1	Regional Centers	GRAND TOTAL
Processing (KSI95)	1,727.0	832.0	4,974.0	7,533.0
Disk (PB)	1.2	1.2	8.7	11.1
Magnet Tape (PB)	16.3	1.2	20.3	37.8

\*Ref: LHC Computing Review report Feb 2001

# Sizes and Scales

- **Storage and computing requirements exceed the capabilities of a single geographic location**
- **10 million GB of data per year ~ 20 million CD ROMs**
- **7500 KSi95 corresponds to PC farms that would cover about 4 acres**
- **How big is 4 acres?**



Common Solutions Group  
September 19, 2002

Research Computing and Grid Technology:  
How These Developments May Impact Higher Education  
I/T Services  
Patrick Dreher, MIT

09/02 ID100-18

# **Astronomy: The National Virtual Observatory**

# National Virtual Observatory

- **Research project linking archival astronomy data sets at geographically dispersed locations and whose goal is to combine these data sets for analysis, comparison and cross correlations of observations**

# National Virtual Observatory

(cont'd)

- **Data sets are geographically distributed and may include space and ground-based observatories, catalogs of multi-wavelength surveys, etc.**
- **Key issue is to interconnect the clusters at each site for**
  - **Data mining**
  - **Sophisticated pattern recognition**
  - **Large-scale statistical correlations**
  - **Discovery of rare objects and temporal variations**

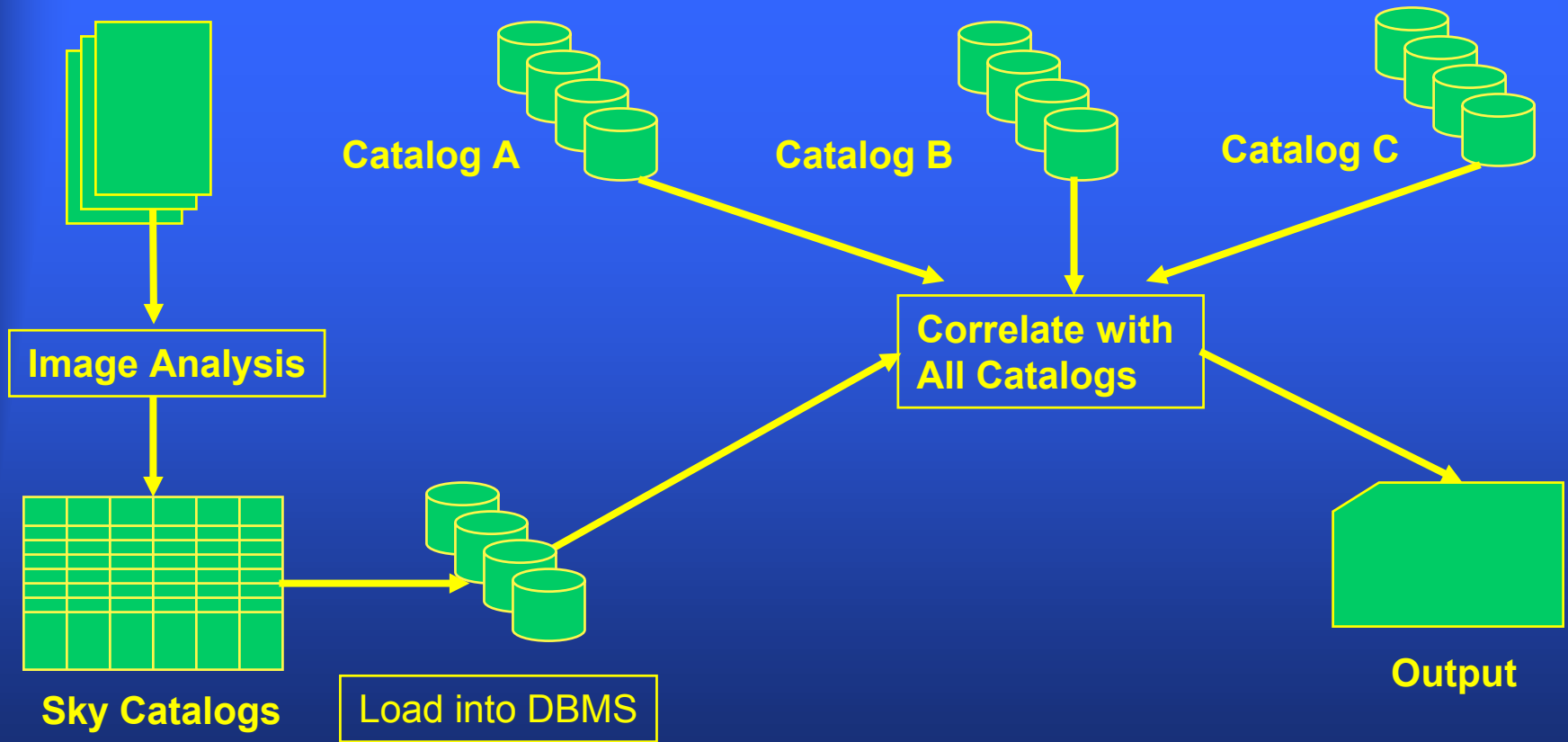
# National Virtual Observatory

(cont'd)

- **This science application requires data management systems capable of utilizing**
  - Data grids
  - Digital libraries
  - Persistent archives
- **These systems will manage**
  - Global scheduling of jobs and positioning of data sets
  - cooperative computing at remote sites (data mining and large-scale statistical correlations components of the science)

# Digital Sky Projects

## National Virtual Observatory (NVO)



# Theoretical Physics Calculations

# **The Strong Force**

## **Understanding the Physics of One of The Four Fundamental Forces of Nature**

# General Physics Goals of Lattice QCD

- Quantum Chromodynamics (QCD) is a theory that has been proposed to explain the physics of the strong interactions – one of the four fundamental forces in the universe
- The overall goal is to  
*understand the rich and complex structure of strongly interacting nuclear matter*

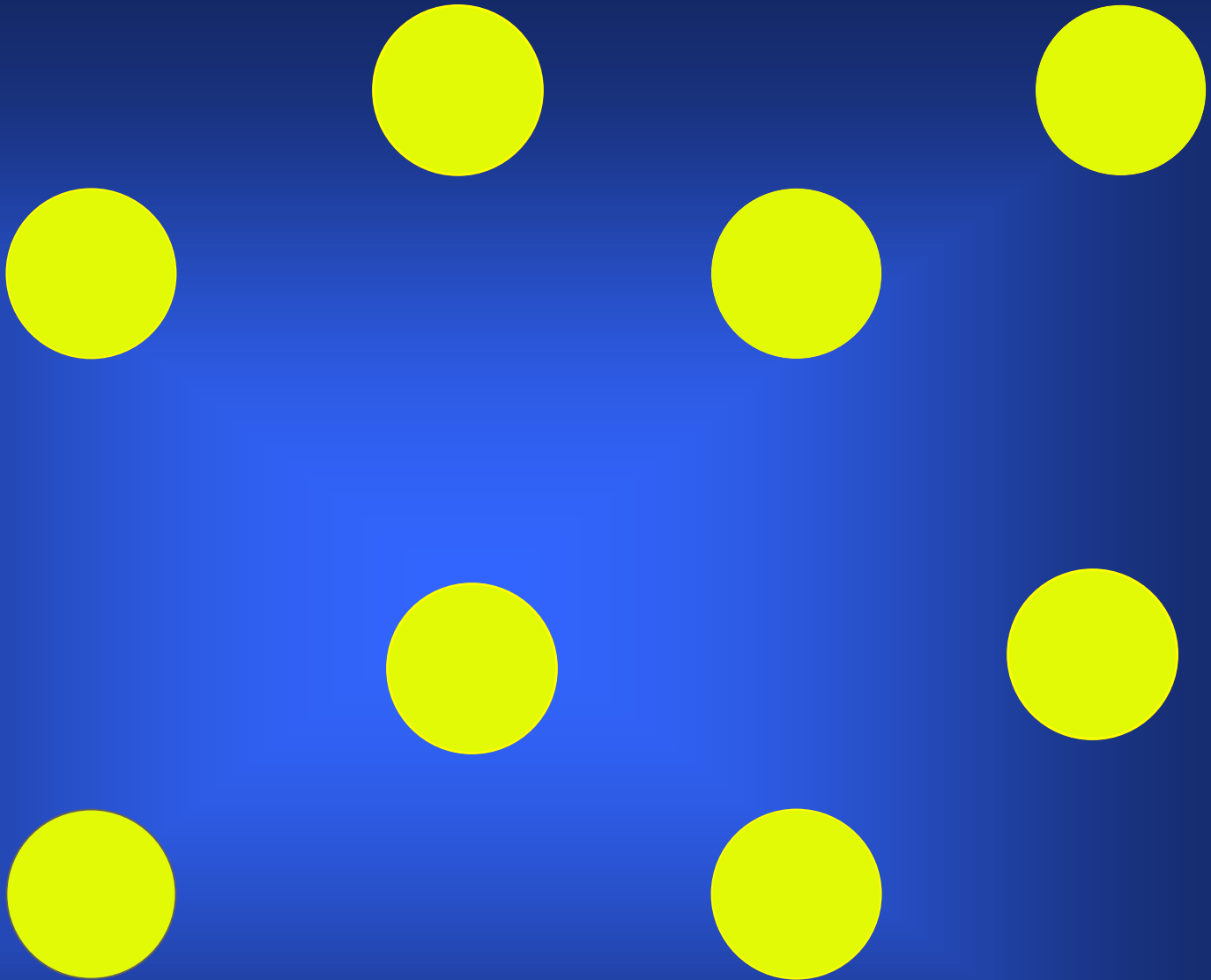
# Need for Computational Resources in Lattice QCD

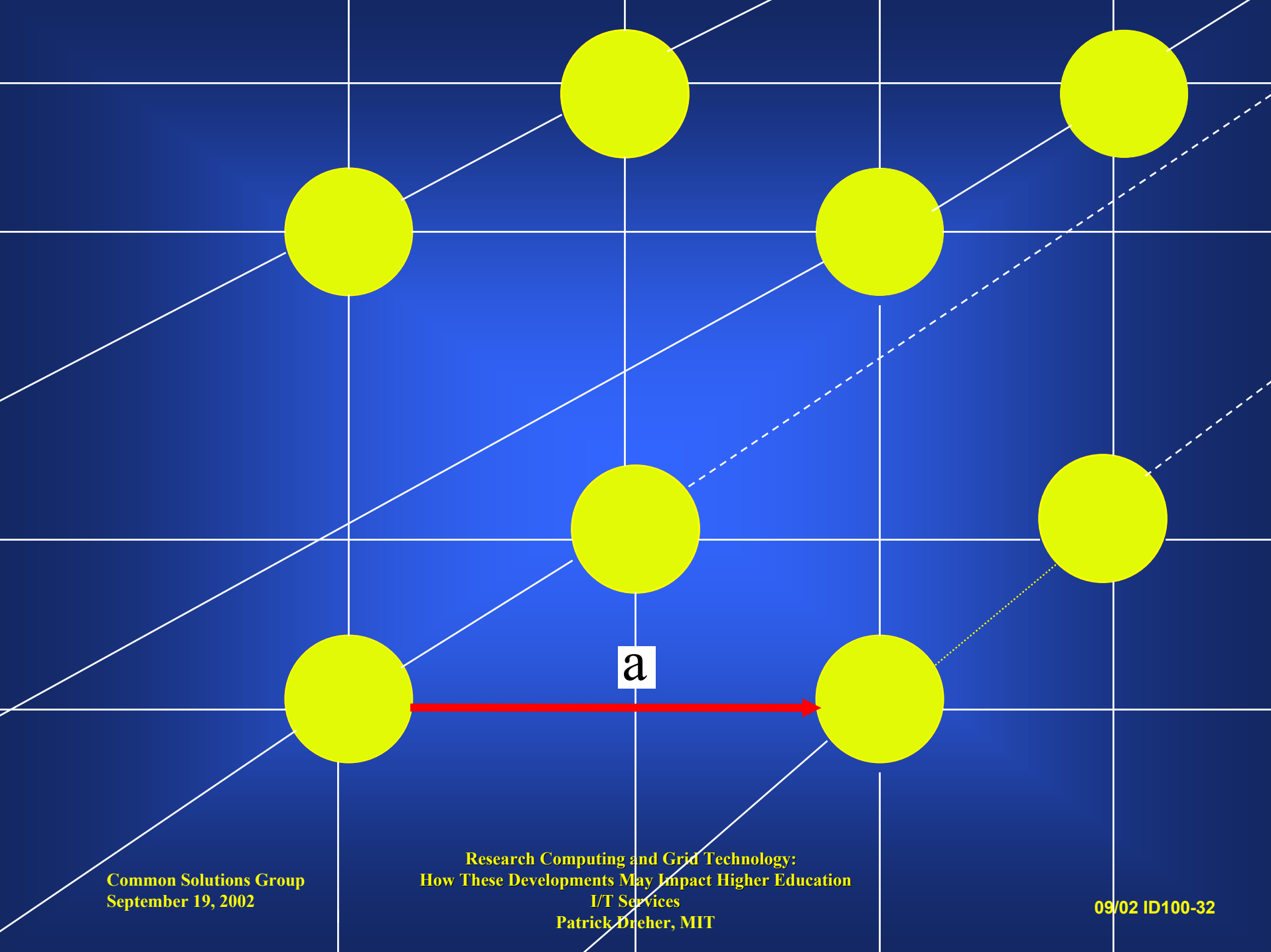
- Traditional theoretical physics analytical techniques used successfully in other theories cannot fully calculate the properties and physics of the strong interactions
- Numerical simulations (lattice QCD) are the only method available to extract the physics of QCD in many regions of the theory

# Why Do Theorists Need Large Amounts of Computer Resources to Study Quantum Chromodynamics?

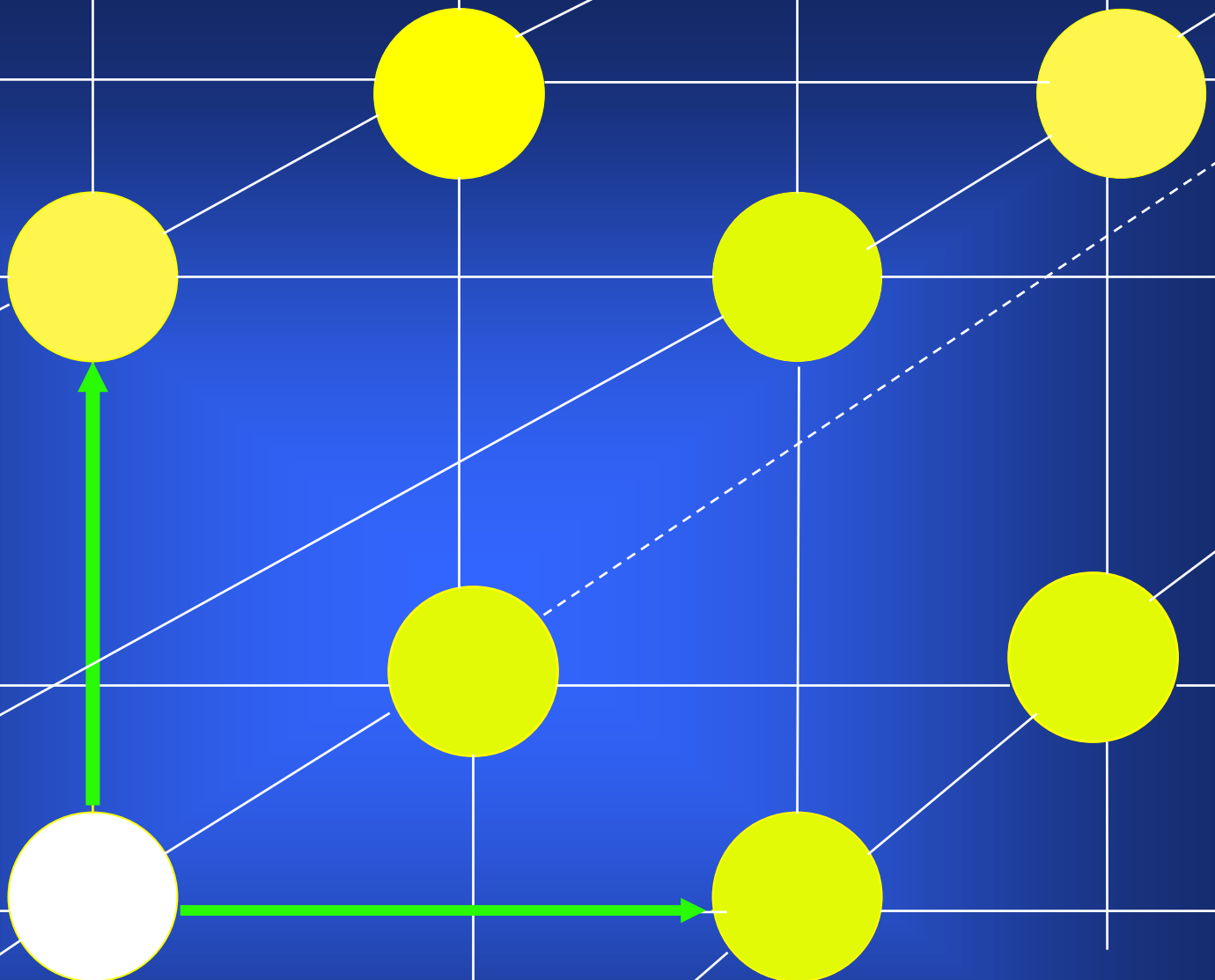






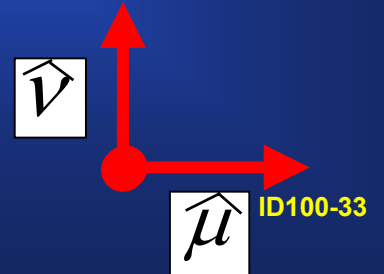


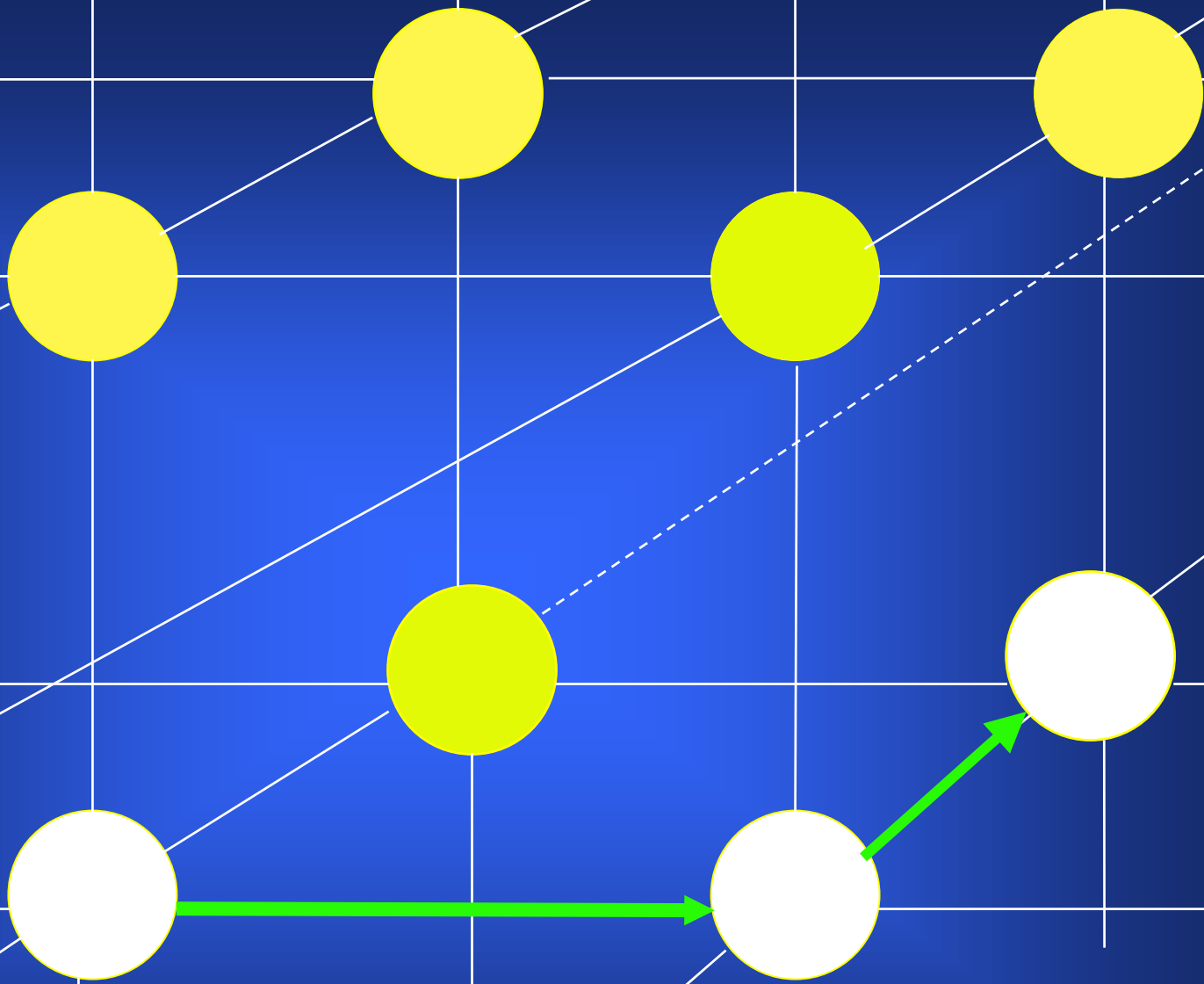
a



Common Solutions Group  
September 19, 2002

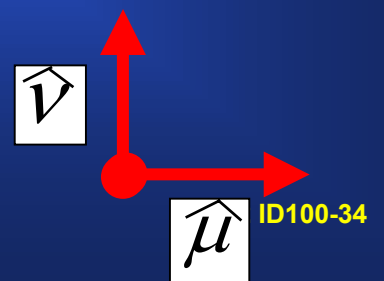
Research Computing and Grid Technology:  
How These Developments May Impact Higher Education  
I/T Services  
Patrick Dreher, MIT

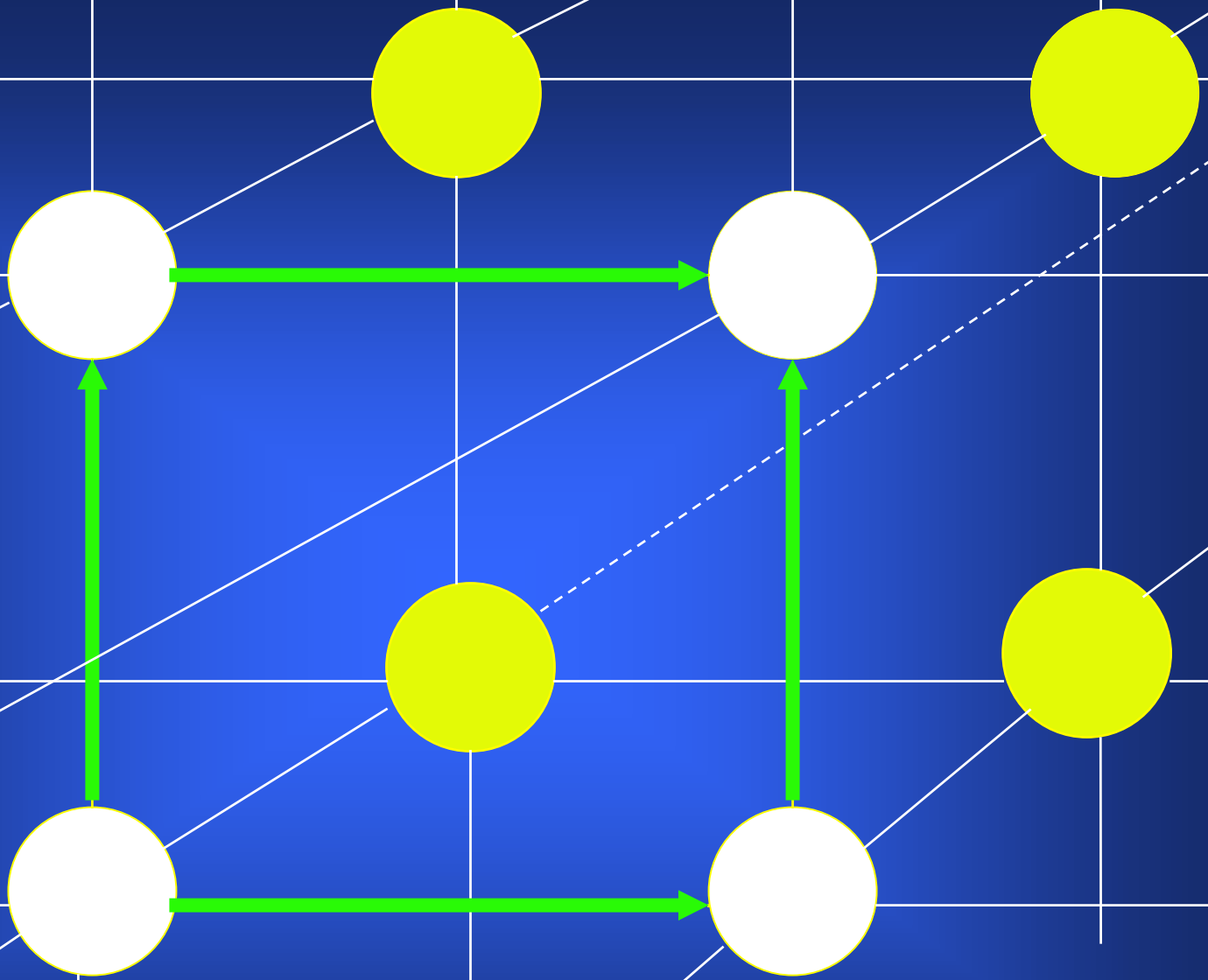




Common Solutions Group  
September 19, 2002

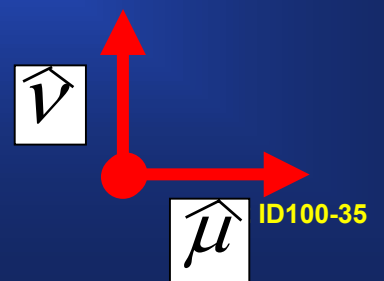
Research Computing and Grid Technology:  
How These Developments May Impact Higher Education  
I/T Services  
Patrick Dreher, MIT

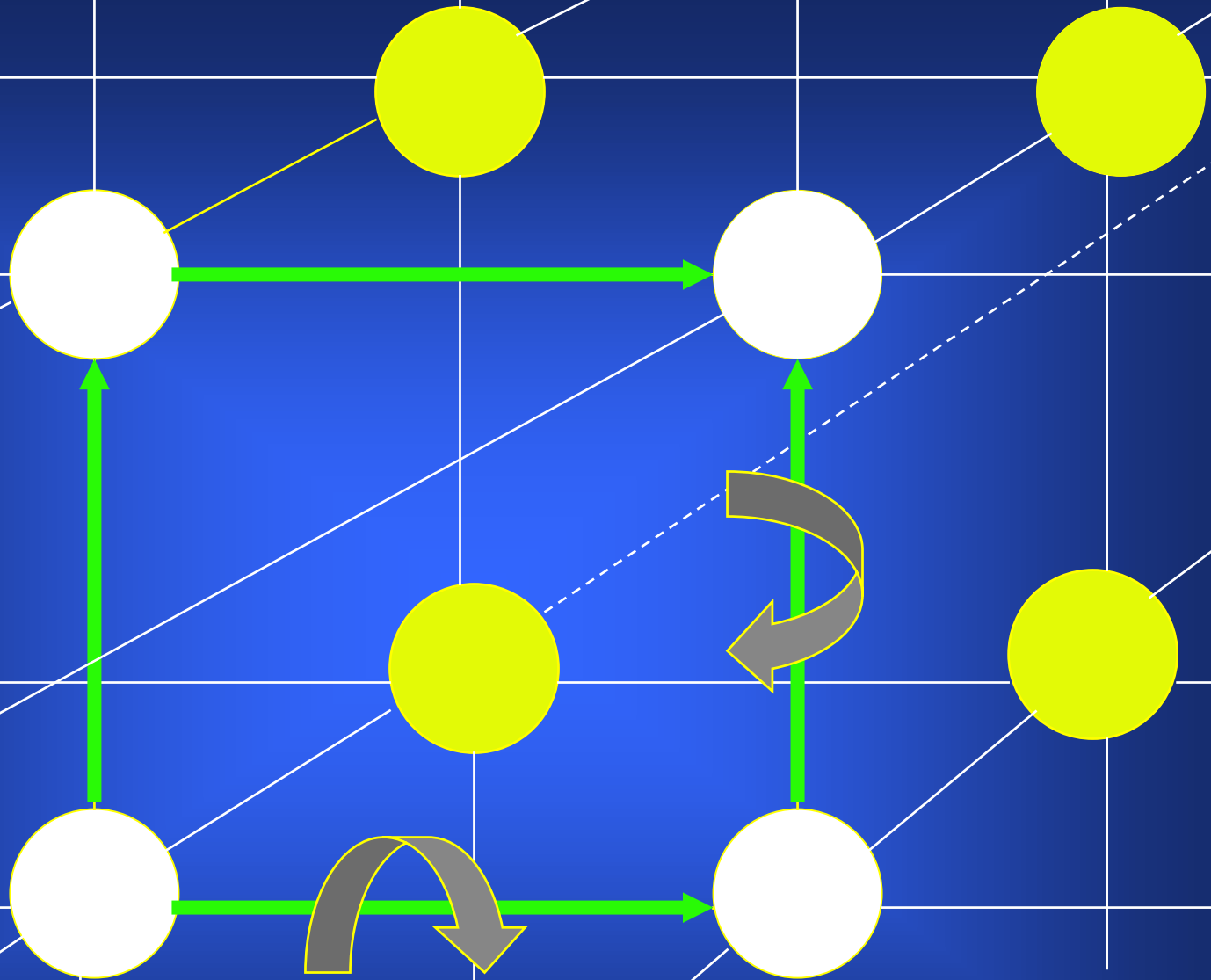




Common Solutions Group  
September 19, 2002

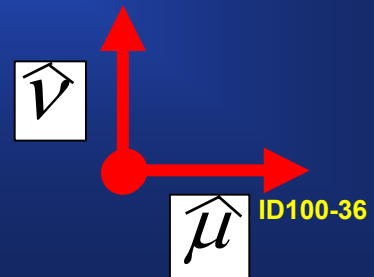
Research Computing and Grid Technology:  
How These Developments May Impact Higher Education  
I/T Services  
Patrick Dreher, MIT

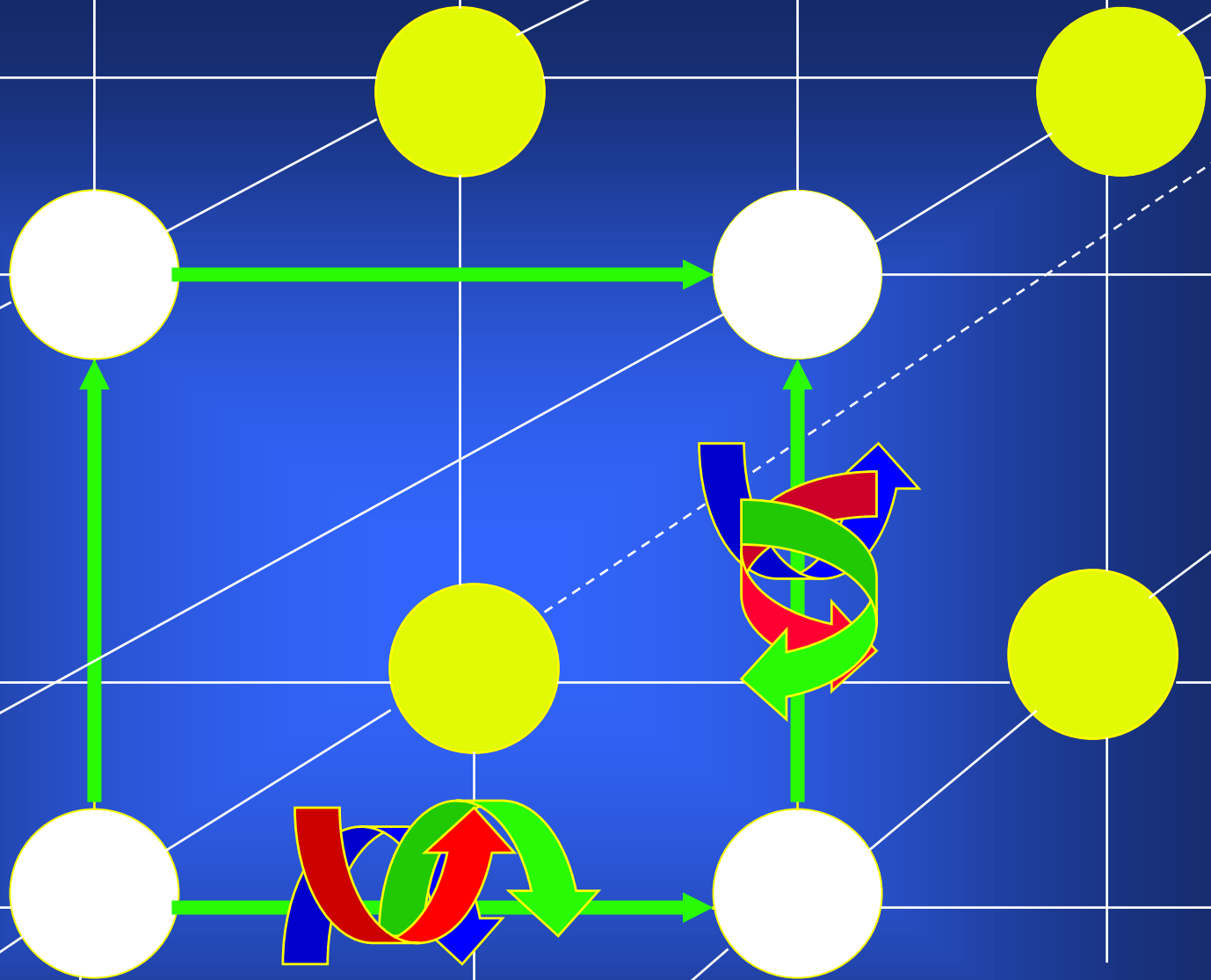




Common Solutions Group  
September 19, 2002

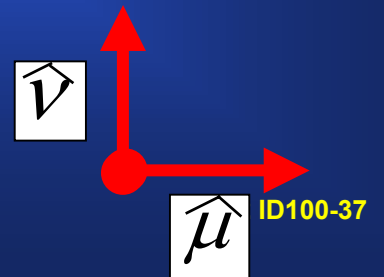
Research Computing and Grid Technology:  
How These Developments May Impact Higher Education  
I/T Services  
Patrick Dreher, MIT

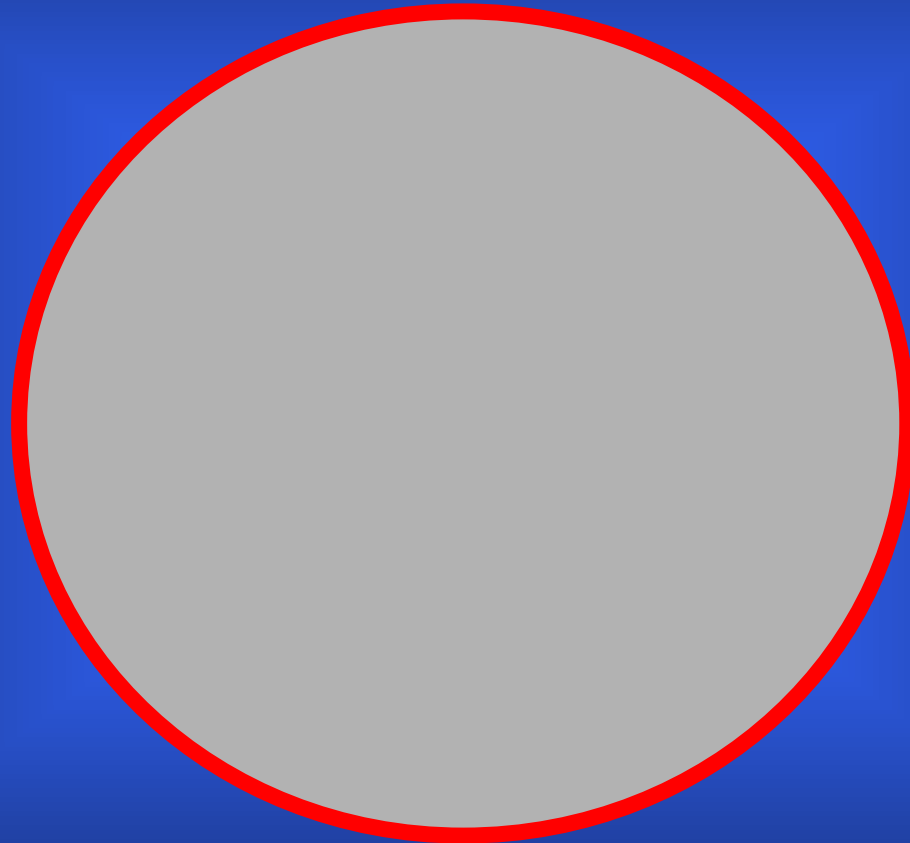


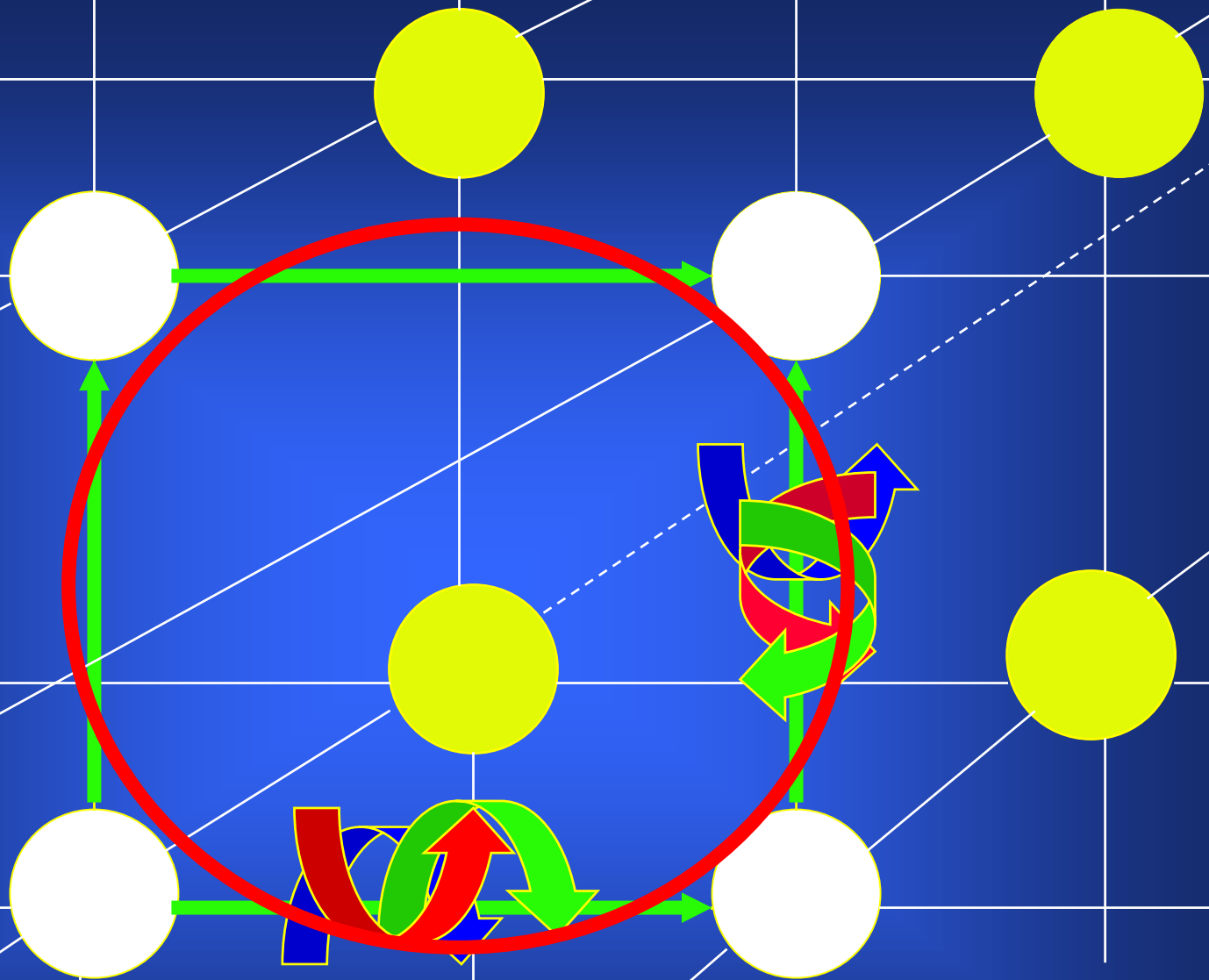


Common Solutions Group  
September 19, 2002

Research Computing and Grid Technology:  
How These Developments May Impact Higher Education  
I/T Services  
Patrick Dreher, MIT





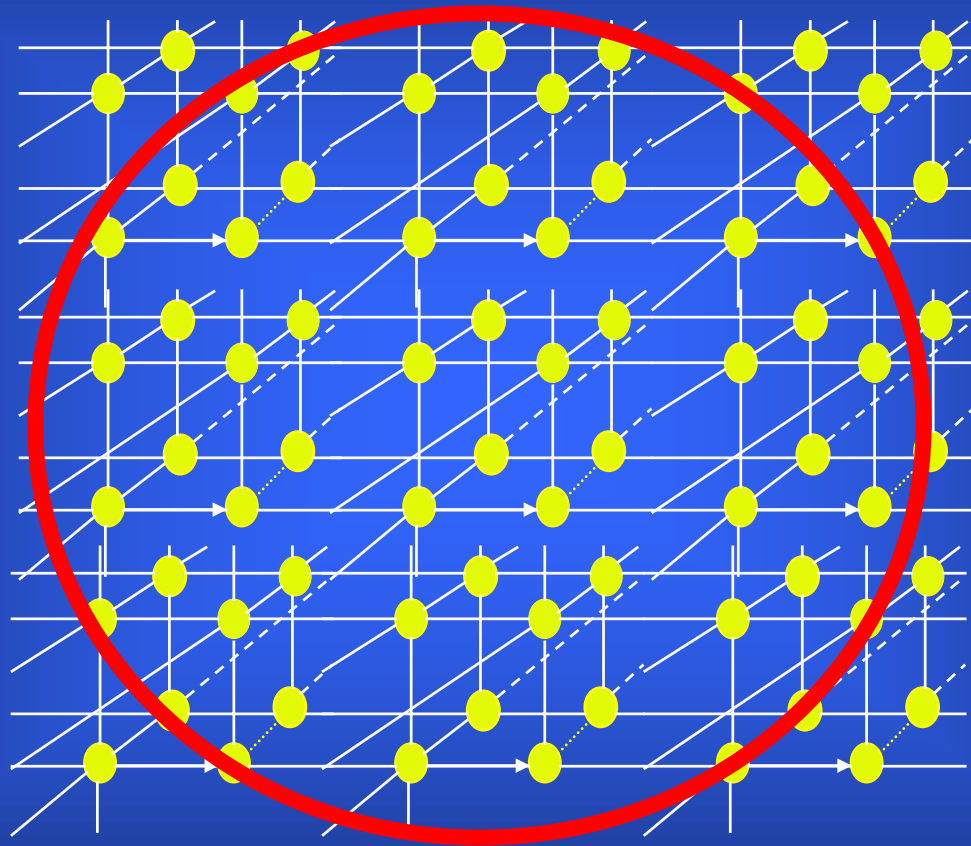


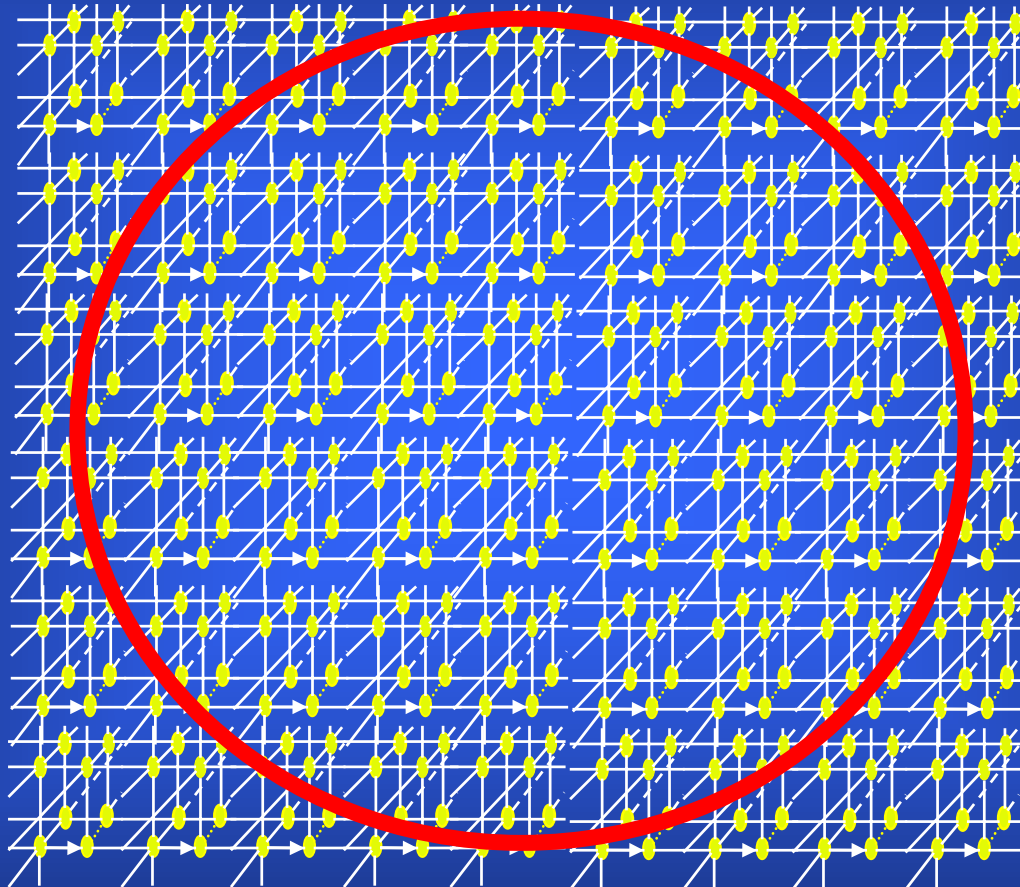
Common Solutions Group  
September 19, 2002

Research Computing and Grid Technology:  
How These Developments May Impact Higher Education  
I/T Services  
Patrick Dreher, MIT



ID100-39





# Time in years for 200 Configurations @ 20 Gflops Sustained \*

Lattice Size  
Time (yr)

.1 fermi (28 <sup>3</sup> )x56	.066fermi (42 <sup>3</sup> )x84	.05 fermi (56 <sup>3</sup> )x112	Pion/rho
.4	4.8	36	.5
(28 <sup>3</sup> )x56	(42 <sup>3</sup> )x84	(56 <sup>3</sup> )x112	.4
1.6	24	156	.4
(36 <sup>3</sup> )x72	(52 <sup>3</sup> )x104	(72 <sup>3</sup> )x144	.3
24	344	2188	.3

Experimental Value =.18

FOR MORE INFORMATION...

<http://xxx.lanl.gov>

hep-lat/9811006

Progress in Lattice Gauge Theory, Steve Sharpe

# Horizons for Research Computing in a Grid Environment

- **What is a “Grid”?**
  - A grid is a system that delivers high-throughput on-demand access to distributed assets on a controlled basis.
- **Factors impacting application of grid technology to scientific research projects**
  - Design
  - Computational support capabilities
  - Deployment

# Design Challenges for Grid Computing

- **Design a distributed computing system capable of excellent:**
  - **Data management**
  - **Data storage**
  - **Communications and data integrity**
- **Grid planning layout and LAN Management**
- **Wide-area networking and internet services**
- **Computer and data security**

# Computational Grid Requirements

- **Computational grids require**
  - **Seamless user access**
  - **Seamless view and access to data and files at geographically separated cluster facilities**
  - **Seamless integration of computational capabilities and capacities across sites**
  - **Regional center coordination**
  - **Network planning**

# Requirements for Grid Deployment

- **Must be communication and control mechanisms (Grid middleware) in place to handle**
  - **Scheduling**
  - **Data Management**
  - **Monitoring**
  - **Error Detection & Recovery**

# Key Issues for Applying Grid Technology to Research Computing

- **Must be able to scale-up grid interconnectivity to large distributed research computing systems**
- **Control operational costs of:**
  - Acquisition
  - Maintenance
  - Operation
- **Grid must provide both efficiency & performance**
- **Large equipment base must be fault tolerant**
- **Systems must be usable within the collaborations**
- **Security of the grid hardware, software, and collaboration data must be insured**

# Grid Resource Management

- **Need a resource allocation manager must support robust job submission and management (meta-computing).**
  - **Site autonomy – resources typically owned and operated by different organizations in different administrative domains (different acceptable use, scheduling, security, policies)**
  - **Heterogeneous substrate – different sites use different local resource management systems (different configurations lead to differences in functionality)**

# Grid Resource Management (cont'd)

- **Resource Management – (cont'd)**
  - **Policy extensibility – meta-computing environment must support local differences in management structures w/o changes to installed local codes**
  - **Co-allocation – Applications requiring simultaneous access to resource requirements at several different sites**
  - **Online control – real-time control of resource requirements that dynamically change during execution**

# Grid Operational Requirements

- **Security -- enables secure authentication and communication over an open network including mutual authentication and single sign-on.**
- **Data Management**
  - **Data Transfer - A high-performance, secure, robust data transfer mechanism**
  - **Replica Catalog - A mechanism for maintaining a catalog of dataset replicas**
  - **Replica Management -- A mechanism that ties together the replica catalog and data transfer**

# Grid Operational Requirements

- **Packaging Technology**
  - software distribution for various platforms and OS-level configurations
  - maintains software coherence across sites and the cluster nodes at each site
  - A "patch-n-build" mechanism for third-party software that is redistributed with patches that automatically apply on installation

# Standardization for Distributed Operations in a Grid Environment

- There now exists standard “packaged” tutorials and turnkey software installations and updates (ex.)

- NPACI Rocks (developed at SDSC and partners at the University of California, Berkeley) provides turnkey software installation and update for Linux clusters, as well as cluster tools such as the Portable Batch System (PBS), Maui Scheduler, and MPICH for Ethernet and Myrinet.

<http://rocks.npaci.edu/>

- The Open Cluster Group (Bald Guy Software, Dell, IBM, Indiana University, Intel, MSC.Software, NCSA, and ORNL)

<http://www.OpenClusterGroup.org/>

# Standardization for Grid Deployment

- Ability to distribute "pre-configured" software to standardize the configurations on all systems
- General software toolkits under development for various components of grid implementation
  - Globus <http://www.globus.org>
- Many research specific grid projects now exist throughout the world
  - PPDG <http://www.ppdg.net/>
  - EU DataGrid <http://www.eu-datagrid.org/>
  - NorduNet <http://www.funet.fi/>
  - Grid Physics Net (GriPhyN) <http://www.griphyn.org>

# **Impacts of Scientific Grid Technology**

## **Research on University Based I/T**

- **Many educational grid projects have similar needs and requirements to research computing projects**
- **Example common design problems**
  - **Mutual authentication**
  - **Single sign-on**
  - **Authorization**
- **Differences**
  - **Network performance demands**
  - **Size and complexity of the data sets**
  - **Scope of applicability**

# **Impacts of Scientific Grid Technology Research on University Based I/T**

- **Issues surrounding authentication**
  - **Interconnection for authentication of educational and research computing grid projects?**
  - **Will there be different certificate authorities for research projects in each research discipline and/or funded by different agencies and/or countries?**
  - **Complex mutual trust of certificate authorities for educational and research projects?**
  - **Maintenance issues for authentication (and revocation)**

# Impacts of Scientific Grid Technology

## Research on University Based I/T

- Many different grid type projects in the educational and research arenas at the present time
- The Grid is sometimes presented as a “flat” structure where everything talks to everything.
- Computer security issues alone make this a completely unrealistic design in the “real world”
- How to design the grid fabric and integrate the educational and scientific components within the university environment?

# Challenges for Universities Providing Research Computing I/T Services

- **Global grids need standards and stability**
- **Risk of divergence**
  - **The promise of the Grid has been not been oversold but the difficulty of developing the requisite Grid infrastructure has been underestimated (Global Grid Forum)**
    - State of the Grid 2002 Dr. Francine Berman, Director, NPACI and SDSC**

# What's Next

- **Observations – rather than conclusions**
  - **Grid projects have a strong research component as well as an educational presence at the university**
  - **University I/T organizations need consider research as well as academic and administrative computing issues in planning for the future**
  - **Educational projects and research projects need to be aware of each other's activities and communicate and exchange information on plans, developments, and results**
  - **This talk only touched on several physical science projects. There are numerous other academic disciplines just as active with grid projects**
  - **Grid technology and developments in computing hardware have fundamentally altered and expanded the possibilities for researchers to tackle problems that were considered impossible 10 years ago**
  - **University I/T needs to play an important role in delivering a robust scientific computing infrastructure for the 1<sup>st</sup> decade of the 21<sup>st</sup> Century**

