the globus project
www.globus.org

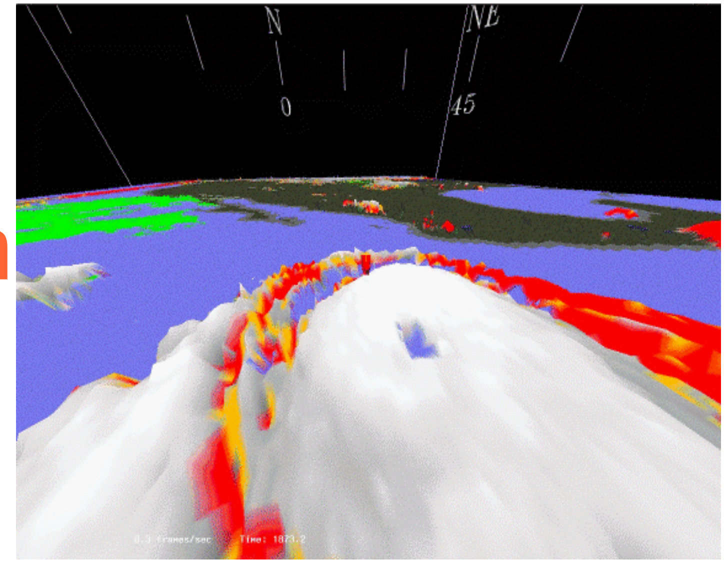
Performance Measurement and Modeling in Computational Grids

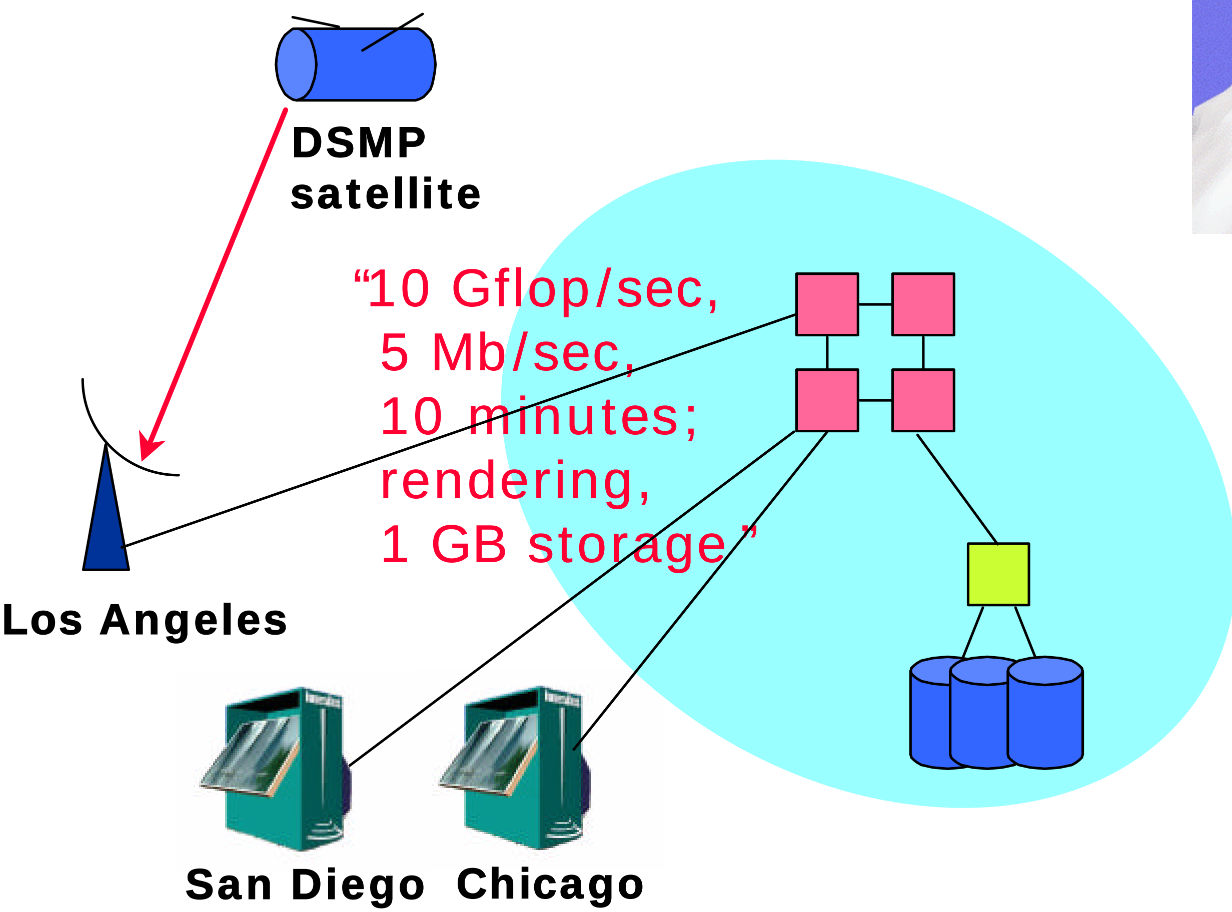
Ian Foster, Noam Freedman,
Brian Toonen
Argonne National Laboratory
University of Chicago

Grid Applications

- High-performance computing in local or wide area networks, for applications that are
  - Distributed by design: e.g., collaborative environments, distributed data analysis, computer-enhanced instruments
  - Distributed by implementation: e.g., metacomputing, high-throughput computing
- Common challenge is to achieve & maintain performance guarantees in heterogeneous, dynamic environments

CNeph: Cloud Detection





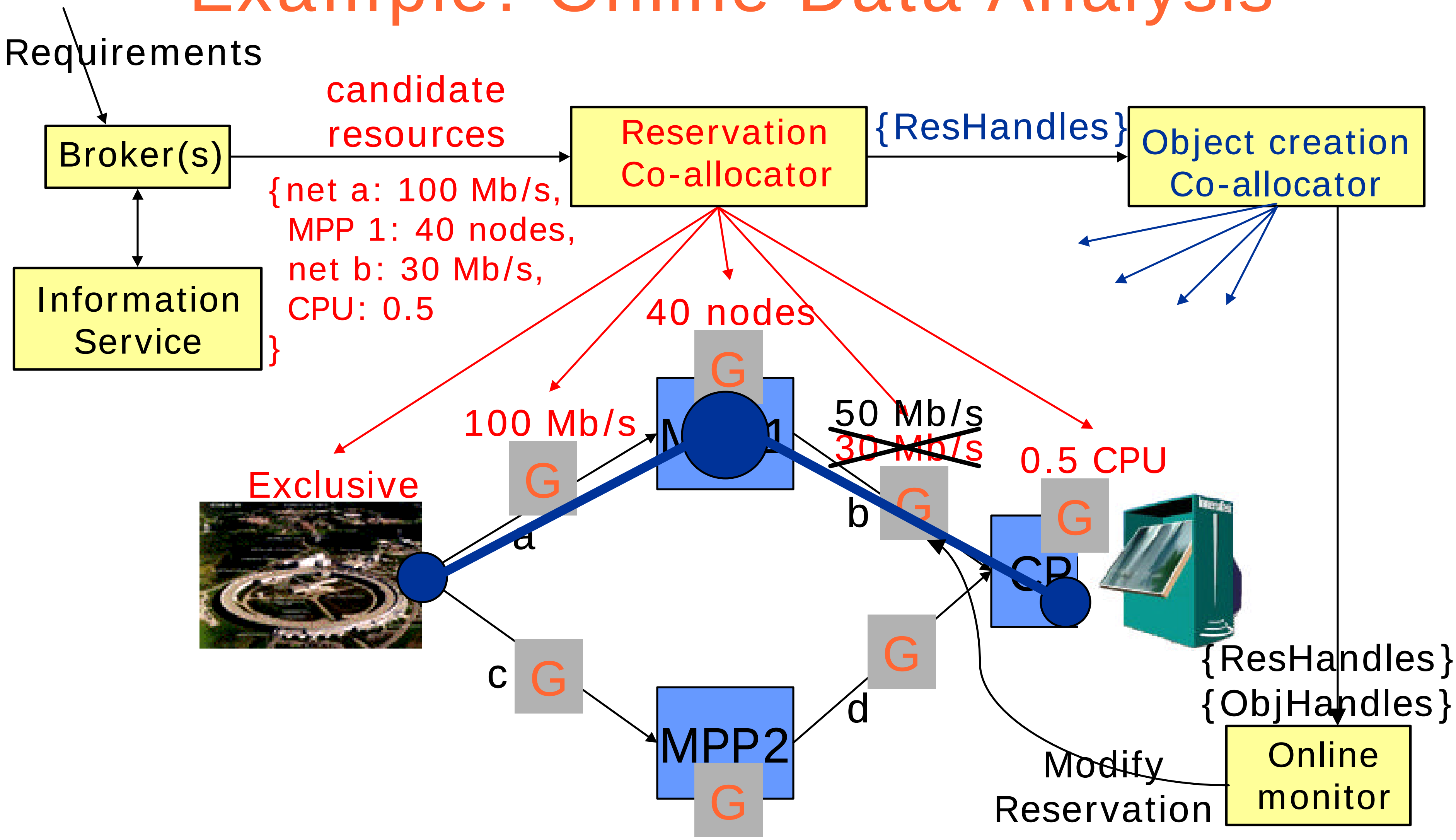
Resource location
↓
Resource allocation
↓
Configuration
↓
Parallel computation
↓
Remote I/O

C. Lee et al., Aerospace Corp.

Performance-Robust Grid Applications

- Increase robustness of grid environment by using implicit or explicit models of application and system performance to
  - Identify resources required to meet application performance requirements
  - Select from among problem specification, algorithm, code variants
  - Establish hierarchical performance contracts
  - Select and manage adaptation strategies when performance contracts are violated

Example: Online Data Analysis



Grids and Delphi, contd.

- Goal: Establish a technology framework for creation of performance-robust applications
- Build on existing technology base
  - Globus toolkit: Grid infrastructure
  - Paradyn/Pablo: Instrumentation, analysis
  - Autopilot: Sensor technology
  - HPC++, MPI: Application programming
- Initial focus is on instrumentation and application studies

Globus Toolkit (Argonne/USC-ISI)

- A set of components providing core services required for grid applications
  - Information, resource management, security, communication, fault detection, data access, etc.
- Used to implement higher-level tools (e.g., MPI, CC++, CAVERN, HPC++, WebFlow, Apples, NEOS, NetSolve) and applications

the globus project
www.globus.org

Using the Globus Toolkit

Chemical engineering | Cosmology | Environmental hydrology | Molecular biology | Scientific instrumentation | Nanomaterials | . . .

Numerical libraries | HPC++ | Condor | Web tools | CAVERN soft | Resource brokers | . . . | Many Worlds | Legion | MPI | SWIG | Nimrod

Accounting | Communications | Information | Scheduling | Fault detection | Security | Instrumentation | QoS | Data access

Convex | SGI | SP | NT | LSF | Kerberos | . . .

the globus project
www.globus.org

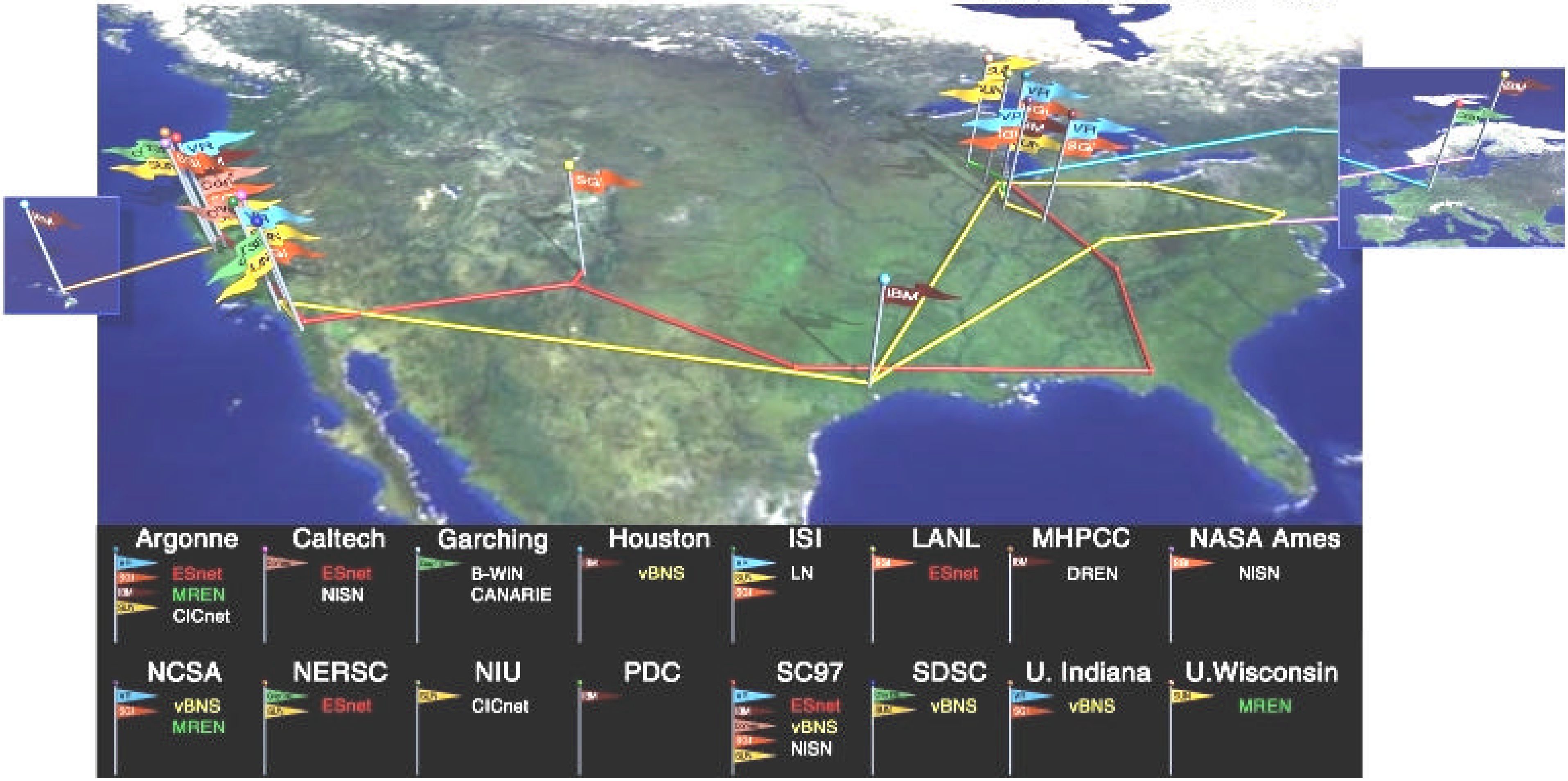
Recent Accomplishments

- Creation of GUSTO computational grid
  - 20+ sites, thousands of processors
- New resource management, security, monitoring, communication technologies
- Largest distributed interactive simulation
  - 100,000 entities on 13 supercomputers
- Release of Globus toolkit v1.0
- International experiments
- 1997 GII "Next Generation" award

the globus project
www.globus.org

GUSTO Computational Grid Testbed

as of November 1997



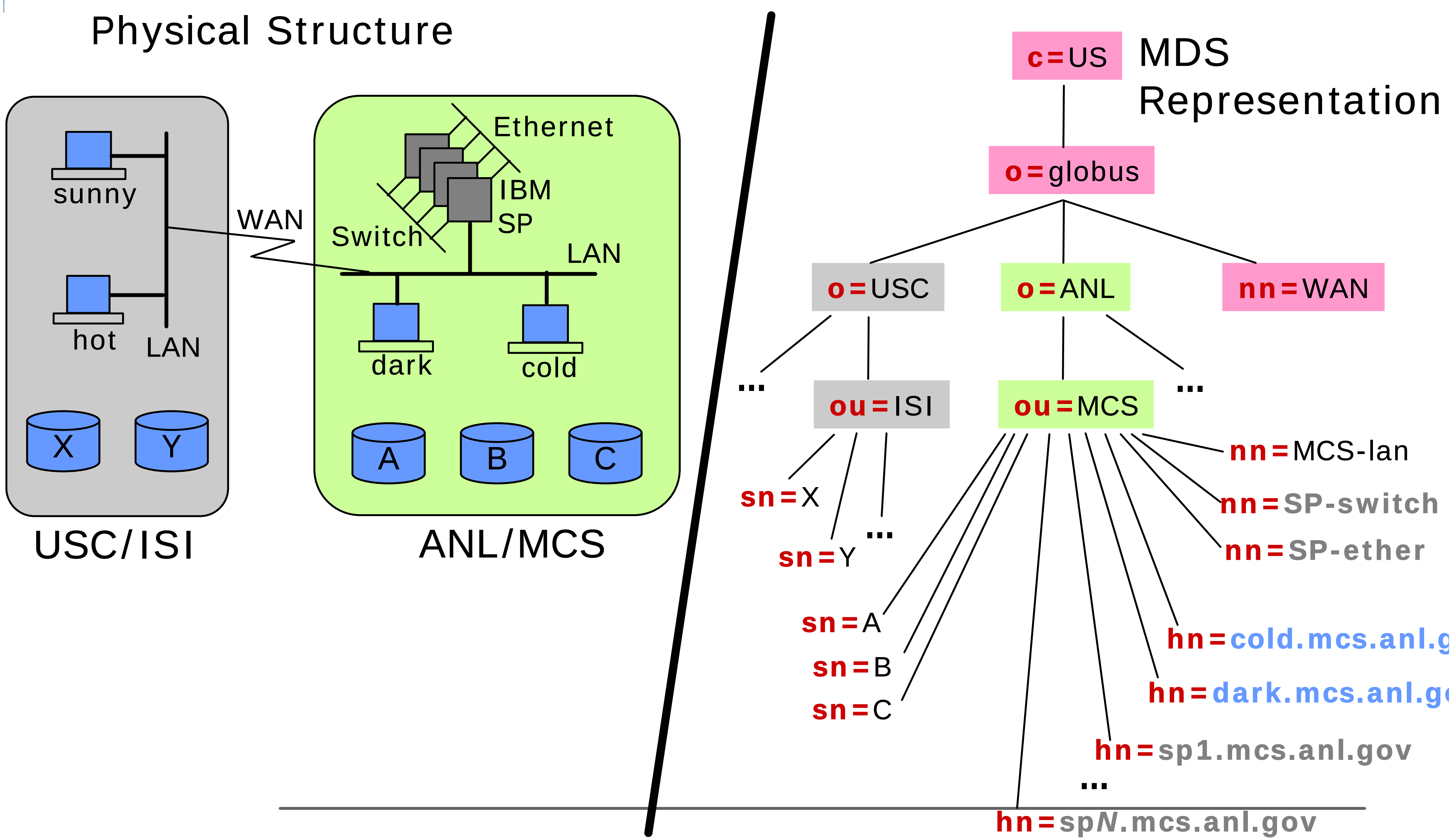
16 sites, 330 computers, 3600 nodes, 2 Teraflop/s, 10 application partners

the globus project
www.globus.org

Delphi and Grids: Initial Steps

- Develop information infrastructure for structural and performance information
- Allow for instrumentation of end-user applications under Globus
  - Paradyn-instrumented applications
- Develop Paradyn instrumentation for the Globus communication library (Nexus)
  - Bandwidth, loss rates, latency, jitter, etc.
- Study performance characteristics of real grid applications

MDS Data Model



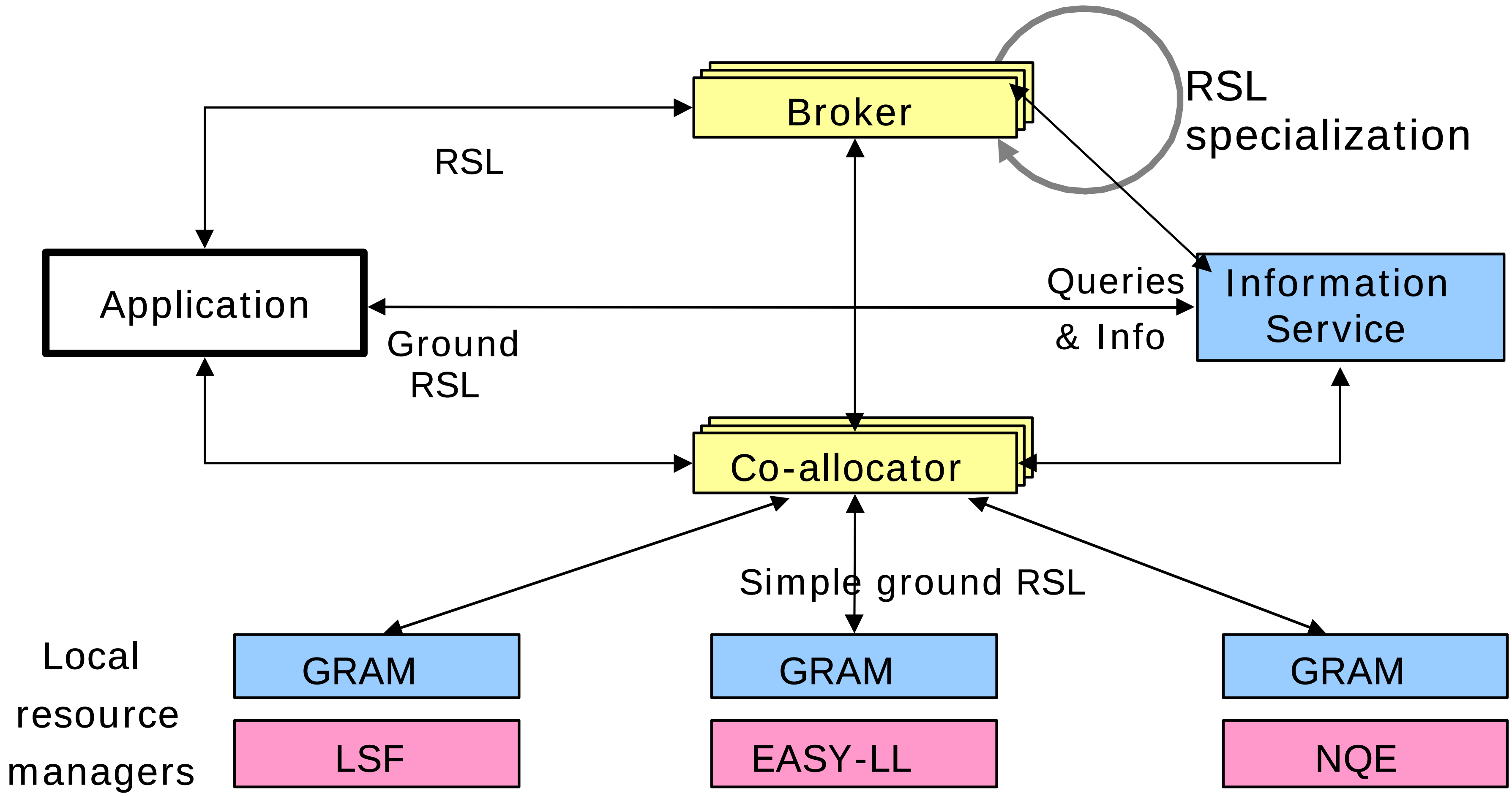
the globus project
www.globus.org

MDS Object Class Example

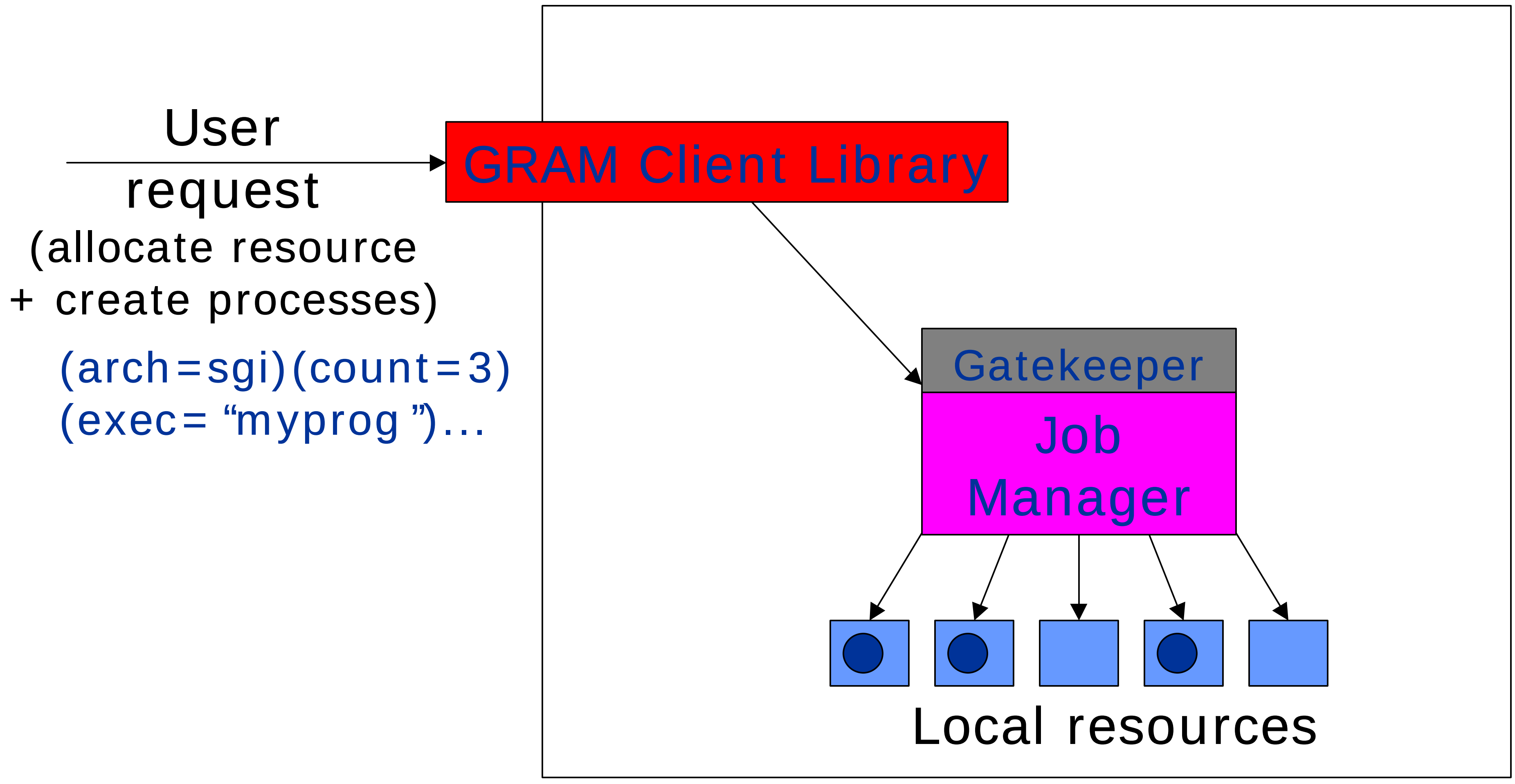
```
GlobusHost OBJECT CLASS
 SUBCLASS OF GlobusResource
  MUST CONTAIN {
   hostName         :: cis,
   type             :: cis,
   vendor           :: cis,
   model            :: cis,
   OStype           :: cis,
   OSversion        :: cis
  }
  MAY CONTAIN {
   networkNode      :: dn,
   totalMemory      :: cis,
   totalSwap        :: cis,
   dataCache        :: cis,
   instructionCache :: cis
  }
```

```
GlobusResource OBJECT CLASS
 SUBCLASS OF top
  MUST CONTAIN {
   administrator :: dn
  }
  MAY CONTAIN {
   manager       :: dn,
   provider      :: dn,
   technician    :: dn,
   description   :: cis,
   documentation :: cis
  }
```

Globus RM Architecture Brokering, Co-allocation, Scheduling, Monitoring



Transparent Instrumentation

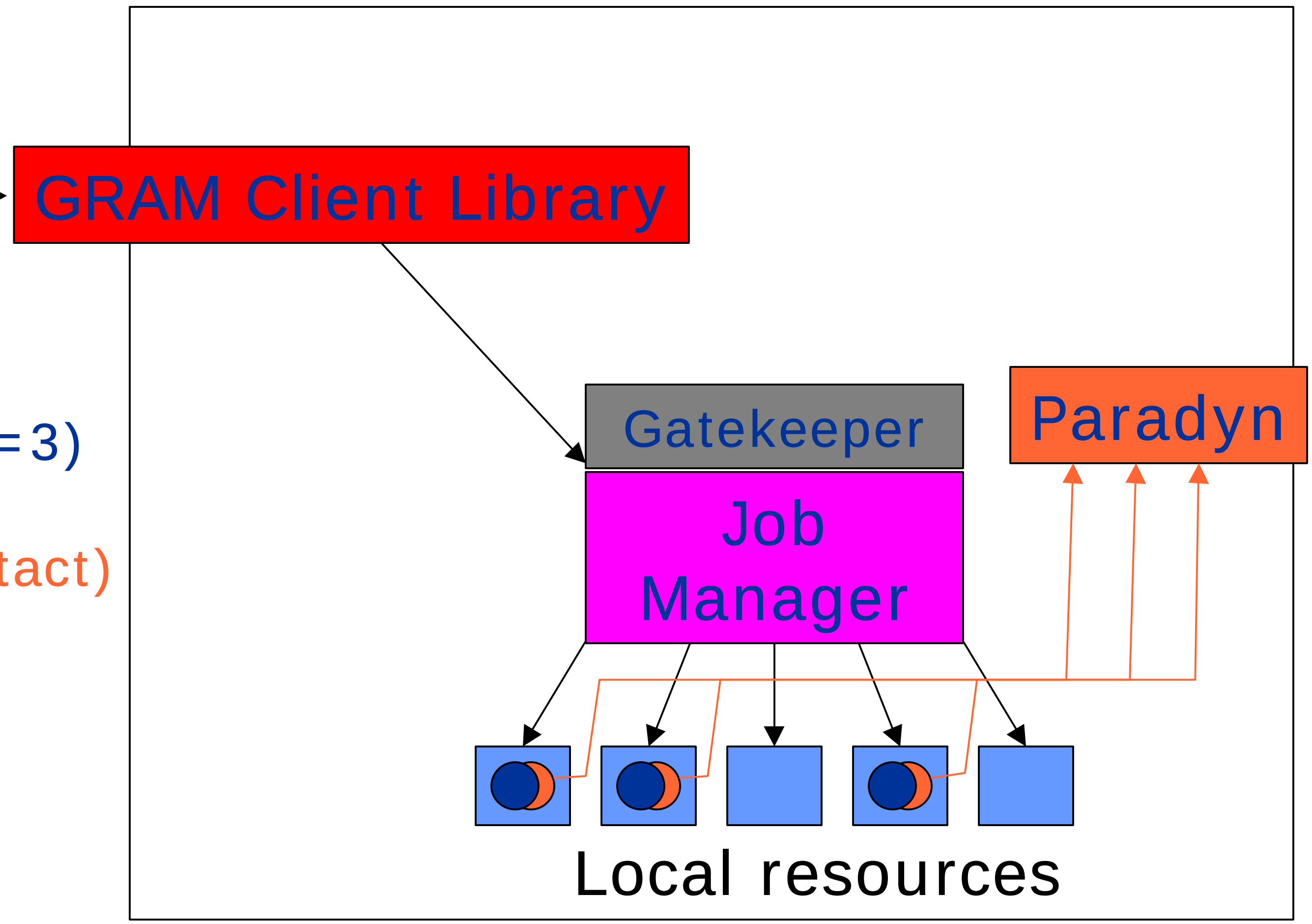


Paradyn Integration

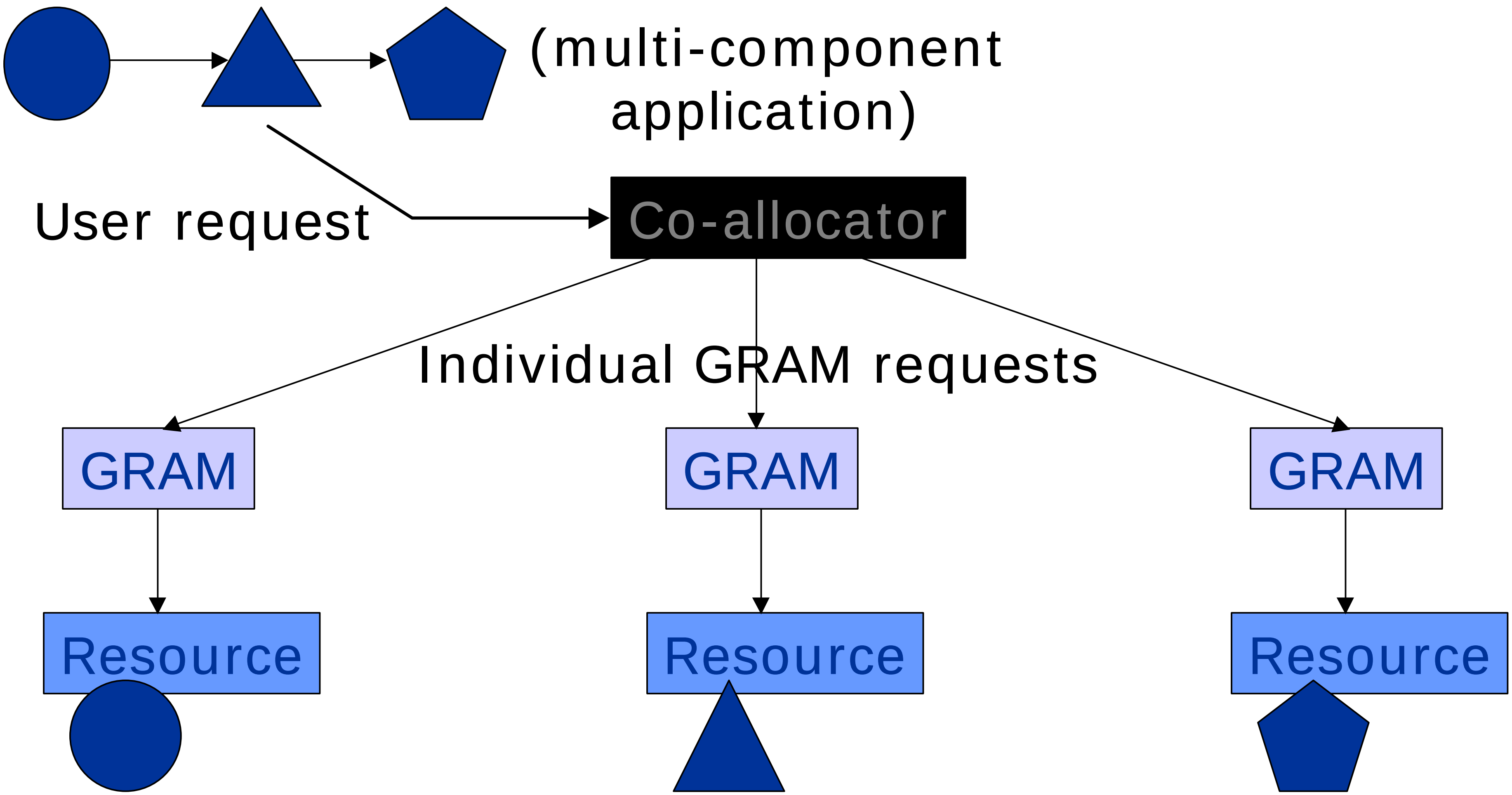
User request

(allocate resource + create processes)

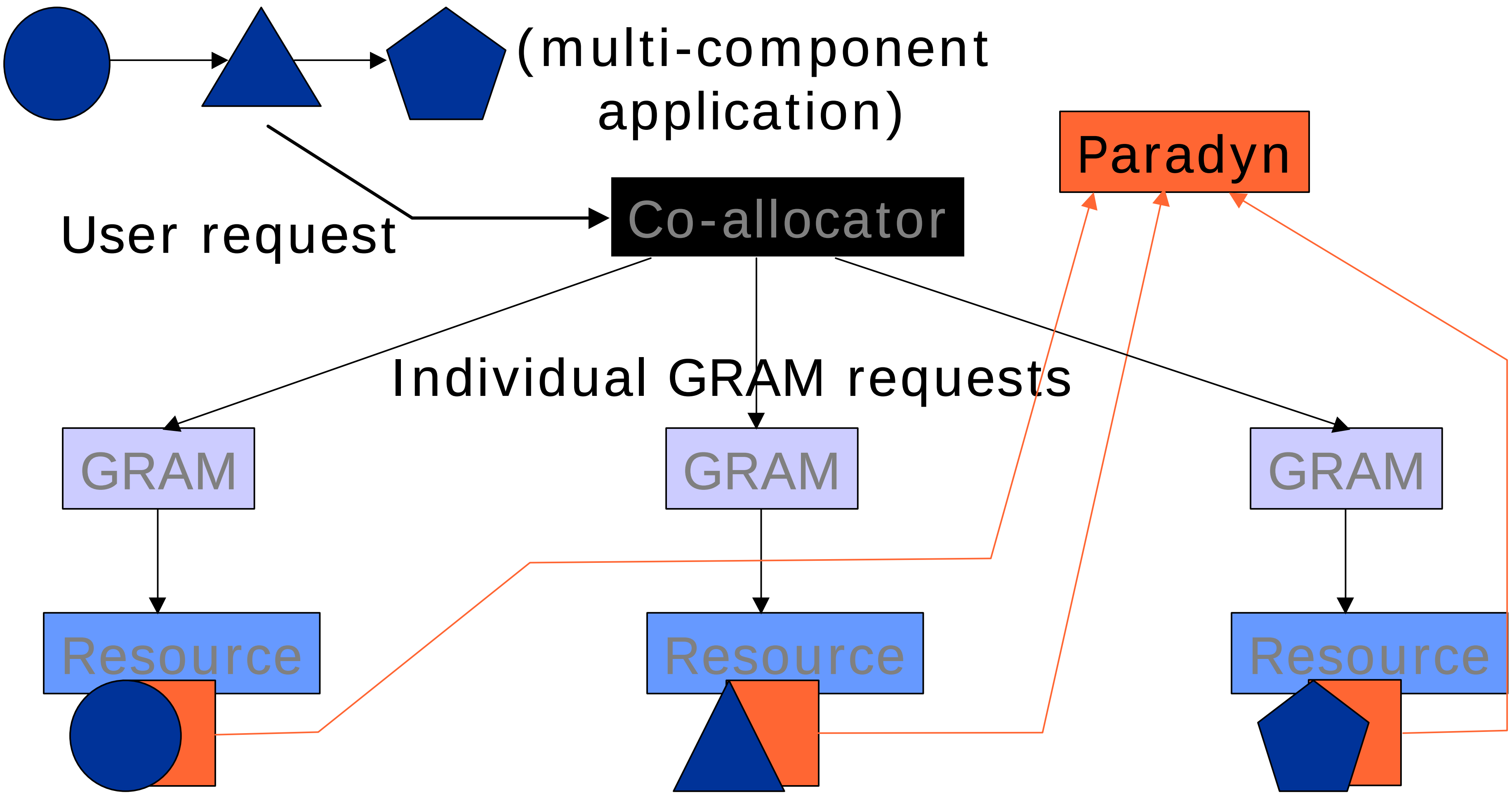
(arch=sgi)(count=3)
(exec= "myprog ")
(paradyn=PDContact)
...



Resource Co-allocation

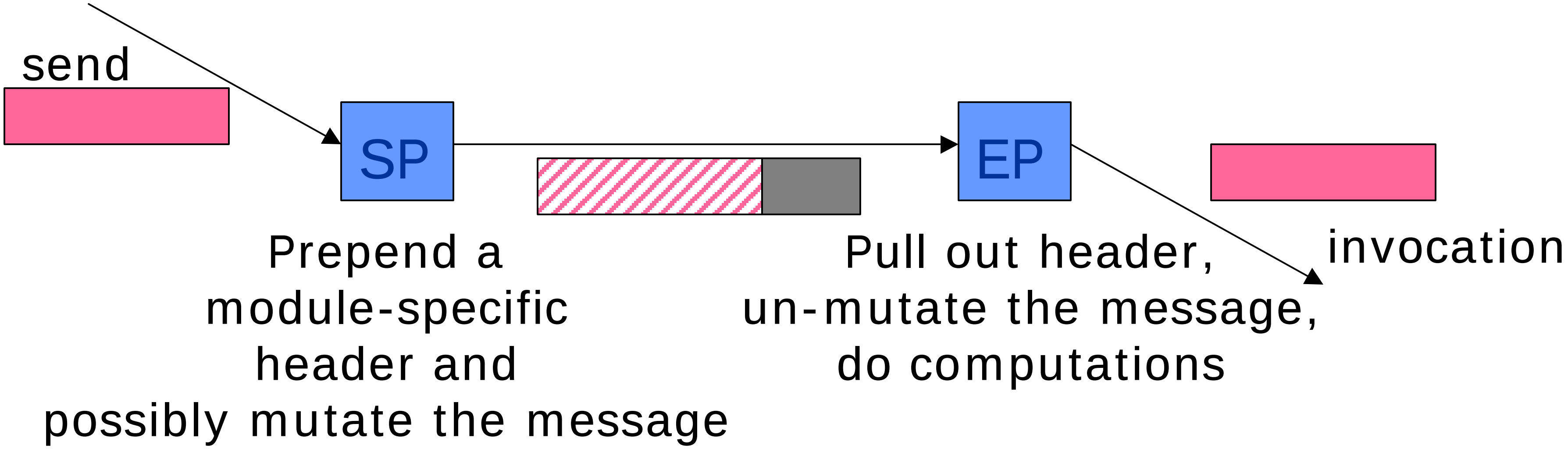


Paradyn Integration



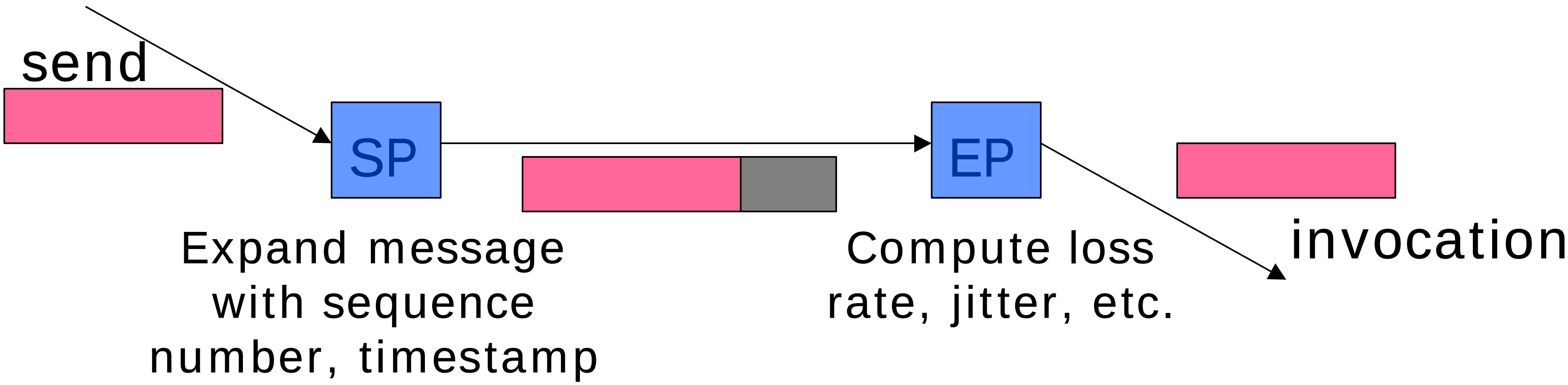
Nexus Transform Modules

- Communication link: startpoints→endpoints
- Comm methods selected on per-CL basis
- User-managed transforms for encryption, compression ... and instrumentation



Nexus Instrumentation

- Insertion of sequence numbers and timestamps for the header
- Computation of loss rates, latency, jittter, etc. on the endpoint.
- Dynamic insertion of transforms?

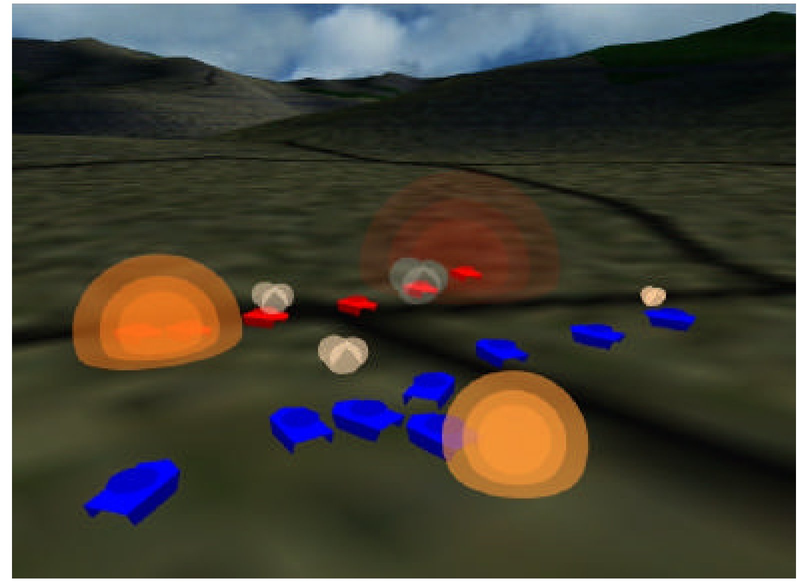


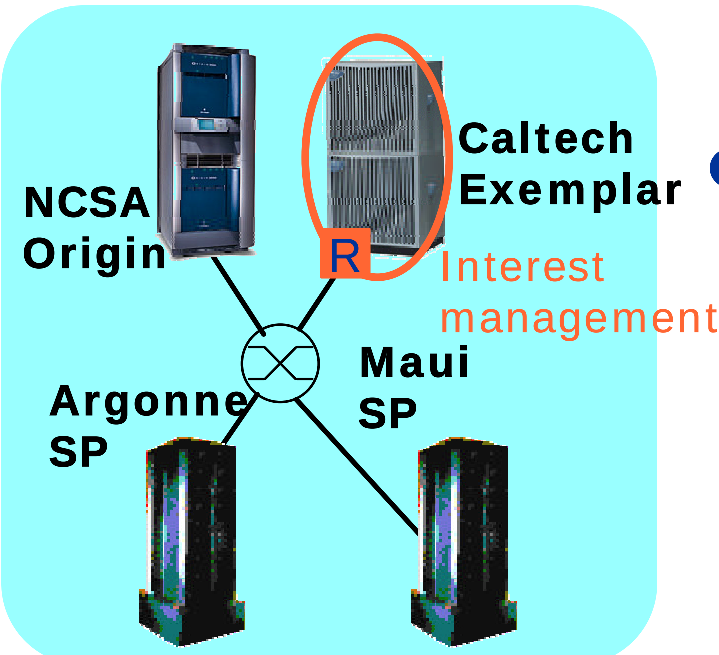
Application Studies

- Distributed computing
  - SF-Express battle-field simulation
  - MPICH-G applications
- Remote I/O
  - RIO remote MPI-I/O library
  - Tardis remote data access library
- Collaborative environments
  - CAVERNsoft collaborative tele-immersion

the globus project
www.globus.org

SF-Express: Distributed Interactive Simulation

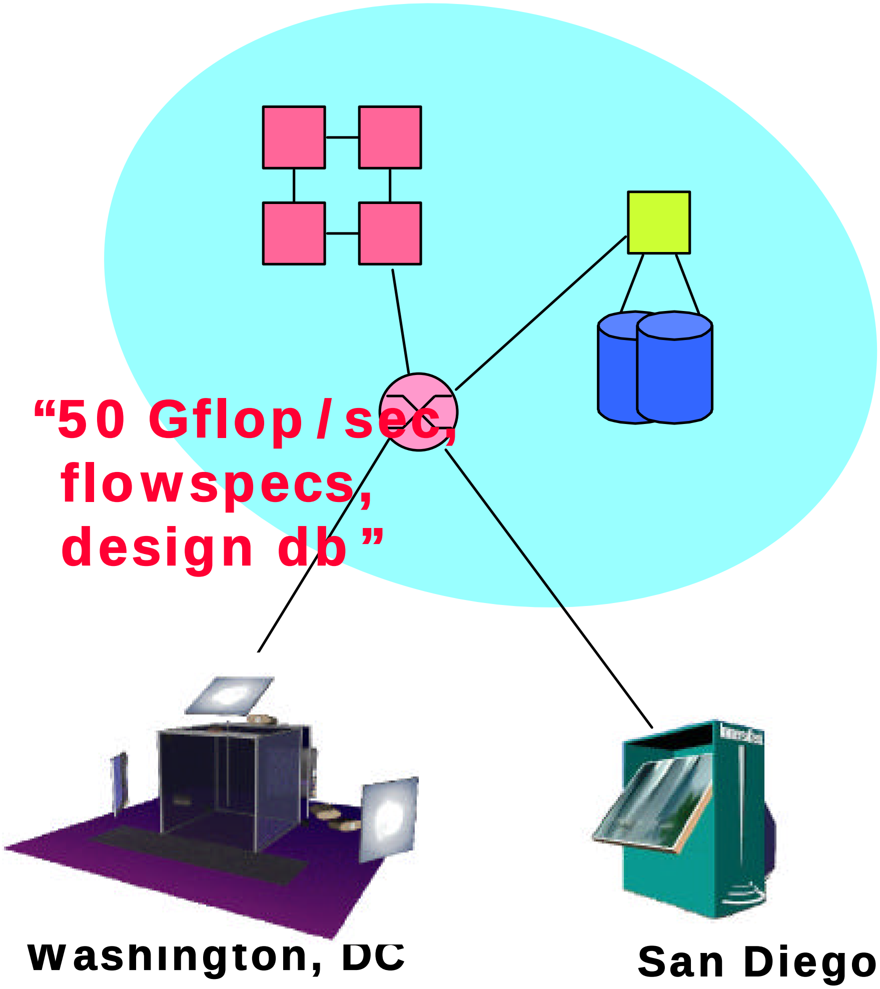




- Performance issues:
  - Organization of computational structure
  - Network requirements
  - Choice of communication methods and protocols
  - Scalability

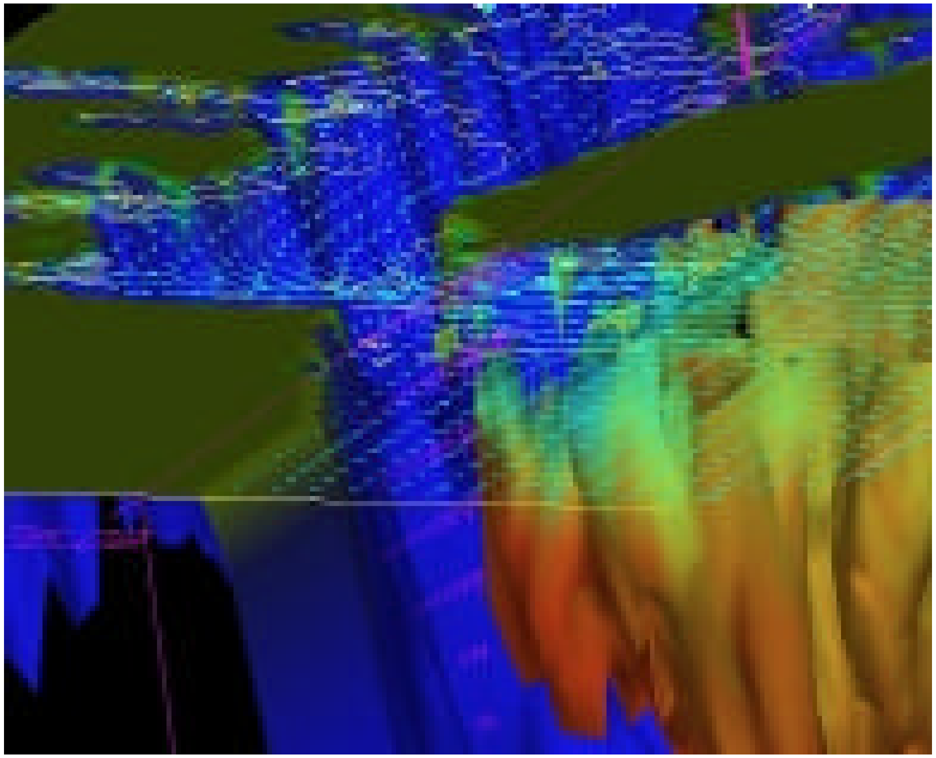
P. Messina et al., Caltech

Application Example: CAVERNsoft Collaborative Design



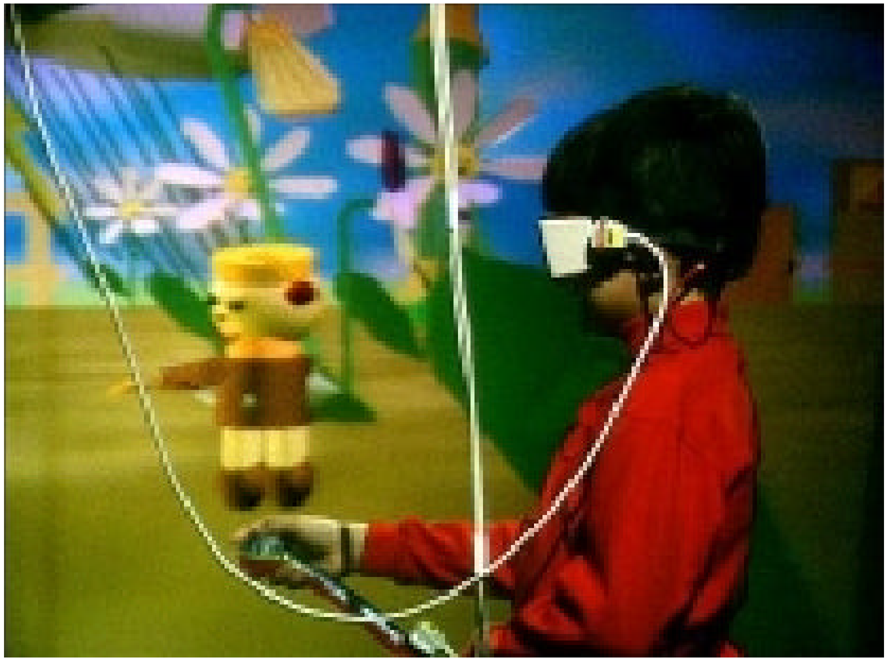
- Allows users to manipulate shared virtual space
- Multiple flows
  - Control, Text, Video, Audio, Database, Simulation, Tracking, Haptics, Rendering
  - 1-10000 Mb/s
- Complex QoS issues

CAVERNsoft Applications



CAVE5D, ODU

NICE, UIC



Images courtesy Jason Leigh, EVL/UIC


Paradyn
Paradyn Main Control
File Setup Phase Visi Help
UIM status : ready
Application name : program: frio_server, machine: (local), user: (self), daem
Data Manager : ready
Processes : PID=16329
Application status : RUNNING
tuva.mcs.anl.gov : application running
RUN PAUSE SAVE EXIT
whereAxis
Selections Navigate Abstraction
Whole Program
Machine
Memory
Process
SyncObject
Code
DEFAULT_MODULE
DYN_MODULE
libc.so.1
libc_psr.so.1
libdyninstRT.so.1
libintl.so.1
libmp.so.1
libnsl.so.1
libsocket.so.1
libw.so.1
t; double-click to expand/un-expand
o expand/un-expand all subtrees of a node
to select/un-select all subtrees of a node
t and move the mouse to scroll freely
rthist
Time Histogram Display
File Actions View
Phase: Global
CPUs bytes/sec
Min:sec
msg bytes recv <Whole Program> (smoothed)
io wait <Whole Program> (smoothed)
io_bytes <Whole Program> (smoothed)
PAN
ZOOM

Next Steps

- Integration with application-level tools
  - HPC++, MPI
- Performance models of individual components
  - See Dennis Gannon's talk
- More detailed application studies
- Experiments in adaptation