The NSF TeraGrid

Mark Bartelt

Center for Advanced Computing Research
California Institute of Technology

mark @ cacr.caltech.edu
http://www.cacr.caltech.edu/~mark
Original Title:

Grid Computing: Hype? Or Buzzword?
The NSF TeraGrid

History
Current Status
Future Directions
History

• PACI (Partnerships for Advanced Computational Infrastructure)
• TCS (Terascale Computing System)
• DTF (Distributed Terascale Facility)
• ETF (Extended Terascale Facility)
PACI Program: NPACI (National Partnership for Advanced Computational Infrastructure)

- San Diego Supercomputer Center (SDSC)
- University of Texas
- University of Michigan
- Caltech
- (others ...
PACI Program: The Alliance (National Computational Science Alliance)

- National Center for Supercomputing Applications (NCSA)
- Argonne National Laboratory
- University of Wisconsin
- Boston University
- University of Tennessee, Knoxville
- University of Kentucky
- Caltech [recently]
- (many, many others ...)

TCS (Terascale Computing System)

- At Pittsburgh Supercomputing Center
- Funded in 2000
- Fully deployed in 2001
- 6 Tflop system (750 quad-processor Alpha nodes)
Distributed TeraScale Facility (DTF)

- Proposal submitted April 2001
- Three-year program
- Four DTF partners:
  - NCSA
  - SDSC
  - Argonne National Laboratory (ANL)
  - Caltech
DTF TeraGrid

- IA64-based Linux clusters at four sites
- Myrinet for intra-cluster connections
- High-bandwidth inter-site interconnect (10 Gbit between every pair of sites)
- Lots of storage
- Grid services based on Globus
Future TeraGrid Authentication Mechanism?
Goals (Measures of Success)

- **New Science**
  - Provide New Capabilities through:
    - Site capabilities that are more powerful than existing PACI resources
    - Combine site resources into a coordinated system
  - To enable:
    - Existing PACI users to deepen their science
    - New users - problems not feasible with today’s PACI resources, require a grid

- **Build an Extensible Grid**
  - Design principles assume heterogeneity and >4 sites
    - A Grid hierarchy similar to Internet hierarchy
      - multiple types, with smaller number of “tightly coupled” and large number of “loosely coupled”
    - Can be **grown**, can be **replicated**, multiple copies can be **combined**
  - Formally documented design: protocols and specifications
    - “Implement this protocol” rather than “Install this magic software”
    - Leverage Global Grid Forum for technical input and dissemination

- **Provide a Pathway for Current Users**
  - Support evolutionary path
    - migration to linux clusters, simple “distributed machine room” model
  - Provide examples, tools, training to exploit grid capabilities
  - User support, user support, and user support
DTF Teragrid: Goals

- Free computational scientists from the “tyranny of distance”
- Seed future cyberinfrastructure
The Arpanet (1969)

DEC 1969

THE ARPA NETWORK

4 NODES
Arpanet (1971)
The Internet (1999)
So ... What was planned?

- IBM Linux clusters
  - open source software and community
- Intel/HP Itanium Processor Family™ nodes
  - “McKinley” processors for commodity leverage
- Very high-speed network backbone
  - bandwidth for rich interaction and tight coupling
- Large-scale storage systems
  - hundreds of terabytes for secondary storage
- Grid middleware
  - Globus, data management, ...
- Next-generation applications
  - breakthrough versions of today’s applications
  - But also, reaching beyond “traditional” supercomputing
DTF Network Topology

- Full N-way mesh
- OC192 links between each pair of sites
The TeraGrid Backbone
So ... What was planned?

**Caltech:** Data collection analysis

**SDSC:** Data-Intensive

**NCSA:** Compute-Intensive

**ANL:** Visualization
Extended Terascale Facility (ETF)

- Proposal submitted June 2002
- New partner (PSC)
- Revised network topology
- Heterogeneity
  - Alpha-based cluster at PSC
  - Power4-based cluster at SDSC
ETF Network Topology

- Major hubs in Los Angeles and Chicago
- 40 Gbit (4 x OC192) connection between hubs
- 3 x OC192 from each DTF site to nearest hub
- Facilitates addition of new sites
ETF TeraGrid

Caltech
- Datawulf IA-32
- 0.5 TF Itanium2
- 90TB

SDSC
- Sun Server
- 7.8 TF Power4
- 1 TF Itanium2
- 300 TB

Federation
- Myrinet
- Fibre Channel

Argonne
- 1.5 TF Itanium2/Madison
- 20 TB

NCSA
- 2 TF Itanium2
- 9.2 TF Madison
- 300 TB

SDSCNCSA
- Myrinet

PSC
- 6TF Alpha EV68
- 1.1 TF Alpha EV7
- 160 TB

Chicago & LA DTF Core Switch/Routers

SDSC
- Myrinet
- Fibre Channel

300TB
- PSC
- Quadrics
- 160 TB
Nostradamus Speaks ...

- The technical challenges will be difficult.
Nostradamus Speaks ...

- The technical challenges will be difficult.
- But the sociopolitical issues will be at least as challenging.
How does it all work?

• Many “working groups”
  – Networking
  – Clusters
  – Performance evaluation
  – Etc. etc. etc ...
TeraGrid Management

Executive Director / Project Manager
Charlie Catlett (UC/ANL)

Chief Architect
Dan Reed (NCSA)

Executive Committee
Fran Berman, SDSC (Chair)
Ian Foster, UC/ANL
Paul Messina, CIT
Dan Reed, NCSA
Rick Stevens, UC/ANL
Charlie Catlett, ANL

Institutional Oversight Committee
Robert Conn, UCSD
Richard Herman UIUC
Dan Meiron, CIT (Chair)
Robert Zimmer, UC/ANL

External Advisory Committee
- Are we enabling new science?
- Are we pioneering the future?

Technical Working Group
- Are we creating an extensible cyberinfrastructure?

User Advisory Committee
- Are we effectively supporting good science?

Policy Oversight
Objectives
Architecture
Implementation

NSF Review Panels

Institutional Oversight Committee
Robert Conn, UCSD
Richard Herman UIUC
Dan Meiron, CIT (Chair)
Robert Zimmer, UC/ANL

Technical Coordination Committee
Project-wide Technical Area Leads
Performance Eval
Brunett (Caltech)
Applications
Williams (Caltech)
Visualization
Papka (ANL)
User Services
Wilkins-Diehr (SDSC)
Towns (NCSA)
Grid Software
Kesselman (ISI)
Butler (NCSA)
Clusters
Pennington (NCSA)
Operations
Sherwin (SDSC)
Data
Baru (SDSC)

Site Coordination Committee
Site Leads
ANL
Evard
CIT
Bartelt
NCSA
Pennington
SDSC
Andrews
PSC
NCAR

NSF MRE Projects
- Internet-2
  McRobbie
- Alliance UAC
  Sugar, Chair
- NPACI UAC
  Kupperman, Chair

NSF Review Panels

External Advisory Committee
- Are we enabling new science?
- Are we pioneering the future?

User Advisory Committee
- Are we effectively supporting good science?

Institutional Oversight Committee
Robert Conn, UCSD
Richard Herman UIUC
Dan Meiron, CIT (Chair)
Robert Zimmer, UC/ANL

Technical Working Group
- Are we creating an extensible cyberinfrastructure?

User Advisory Committee
- Are we effectively supporting good science?
How does it all work?

- Every working group includes people from all TeraGrid sites.
How does it all work?

• Every working group includes people from all TeraGrid sites.
• How the heck do you coordinate all these people?
TeraGrid Management

Executive Committee
Fran Berman, SDSC (Chair)
Ian Foster, UC/ANL
Paul Messina, CIT
Dan Reed, NCSA
Rick Stevens, UC/ANL
Charlie Catlett, ANL

Chief Architect
Dan Reed
(NCSA)

Executive Committee
Fran Berman, SDSC (Chair)
Ian Foster, UC/ANL
Paul Messina, CIT
Dan Reed, NCSA
Rick Stevens, UC/ANL
Charlie Catlett, ANL

Institutional Oversight Committee
Robert Conn, UCSD
Richard Herman UIUC
Dan Meiron, CIT (Chair)
Robert Zimmer, UC/ANL

External Advisory Committee
- Are we enabling new science?
- Are we pioneering the future?

Technical Working Group
- Are we creating an extensible cyberinfrastructure?

Policy Oversight
Objectives
Architecture
Implementation

NSF ACIR
Project Director
Rick Stevens
(UC/ANL)

Chief Architect
Dan Reed
(NCSA)

Executive Committee
Fran Berman, SDSC (Chair)
Ian Foster, UC/ANL
Paul Messina, CIT
Dan Reed, NCSA
Rick Stevens, UC/ANL
Charlie Catlett, ANL

User Advisory Committee
- Are we effectively supporting good science?

External Advisory Committee
- Are we enabling new science?
- Are we pioneering the future?

Institutional Oversight Committee
Robert Conn, UCSD
Richard Herman UIUC
Dan Meiron, CIT (Chair)
Robert Zimmer, UC/ANL

Technical Coordination Committee
Project-wide Technical Area Leads
Performance Eval
Brunett (Caltech)
Applications
Williams (Caltech)
Visualization
Papka (ANL)
User Services
Wilkins-Diehr (SDSC)
Towns (NCSA)
Operations
Sherwin (SDSC)
Data
Baru (SDSC)
Grid Software
Kesselman (ISI)
Butler (NCSA)
Clusters
Pennington (NCSA)
Networks
Winkler (ANL)

Site Coordination Committee
Site Leads
- ANL Evard
- CIT Bartelt
- NCSA Pennington
- SDSC Andrews
- PSC
- NCAR

NSF MRE Projects
Internet-2 McRobbie
Alliance UAC Sugar, Chair
NPACI UAC Kupperman, Chair

NSF Review Panels
Institutional Oversight Committee
- Robert Conn, UCSD
- Richard Herman UIUC
- Dan Meiron, CIT (Chair)
- Robert Zimmer, UC/ANL

External Advisory Committee
- Are we enabling new science?
- Are we pioneering the future?

Technical Working Group
- Are we creating an extensible cyberinfrastructure?
How does it all work?

- Every working group includes people from all TeraGrid sites.
- How the heck do you coordinate all these people?
- We all seem to spend half our lives on conference calls, and the other half replying to e-mail.
How does it all work?

- Every working group includes people from all TeraGrid sites.
- How the heck do you coordinate all these people?
- We all seem to spend half our lives on conference calls, and the other half replying to e-mail.
- The “herding cats” analogy is apt.
Commercial Break
Potholes on the Road to Production

• Original timetable:
  – Sep 2002: Initial delivery of phase-1 systems
  – Mar 2003: Friendly users
  – July 2003: Production

• Hmm, what if there are unexpected problems?
Problem-free? Hah!

• Hardware delivery behind schedule.
  – This should have come as a surprise?
Problem-free? Hah!

- Hardware delivery behind schedule.
  - This should have come as a surprise?
- Gack! Numerical errors!
  - Bug in floating-point software assist code.
Problem-free? Hah!

- Hardware delivery behind schedule.
  - This should have come as a surprise?
- Gack! Numerical errors!
  - Bug in floating-point software assist code.
  - Kernel bug: Floating-point registers sometimes not being saved/restored properly on context switch.
Problem-free? Hah!

- Hardware delivery behind schedule.
  - This should have come as a surprise?
- Gack! Numerical errors!
  - Bug in floating-point software assist code.
  - Kernel bug: Floating-point registers sometimes not being saved/restored properly on context switch.
  - SUPERB support from IBM + Intel on first problem, and from IBM + SuSE on second.
Problem-free? Hah!

- Breaking news; this just in ...
  - Problem with Itanium2 processor (hit the trade press just yesterday)
  - {Intel|IBM} working on a solution
Unsolved (or partially-solved) problems

- Metascheduling
- Coordinated advance reservation
- On-demand computing
- Very large datasets
  - Even at 10 Gbit/second, 100 TBytes takes approximately one day to move
  - Possible solutions: Datacutter and similar tools
High-Bandwidth TeraGrid Transport Mechanism
So ... Where are we?

- Target production date has slipped three months (from beginning of July to beginning of October).
So ... Where are we?

- Target production date has slipped three months (from beginning of July to beginning of October).
- Everybody is overworked and stressed out.
So ... Where are we?

- Target production date has slipped three months (from beginning of July to beginning of October).
- Everybody is overworked and stressed out.
- But ... We’re having LOADS of fun!
Let us close with a prayer ...
Let us close with a prayer …

Anointed with oil,
on troubled waters,
oh heavenly Grid,
help us bear up Thy standard,
our chevron flashing bright
across the gulf of compromise.
Remember: TeraGrid is ...
Questions?

http://www.teragrid.org