Thoughts on HPC Facilities Strategies for DOE Office of Science: Grids to Petaflops

Rick Stevens

Argonne National Laboratory

University of Chicago

Outline

- Update on NSF's distributed terascale facility
- What grid and facilities strategy is appropriate for DOE?
- Limits to cluster based architectures
- New paths to petaflops computing capability
- Grid implications of affordable petaflops
- Summary and recommendations

NSF TeraGrid Approach [$53M in FY01-FY03]

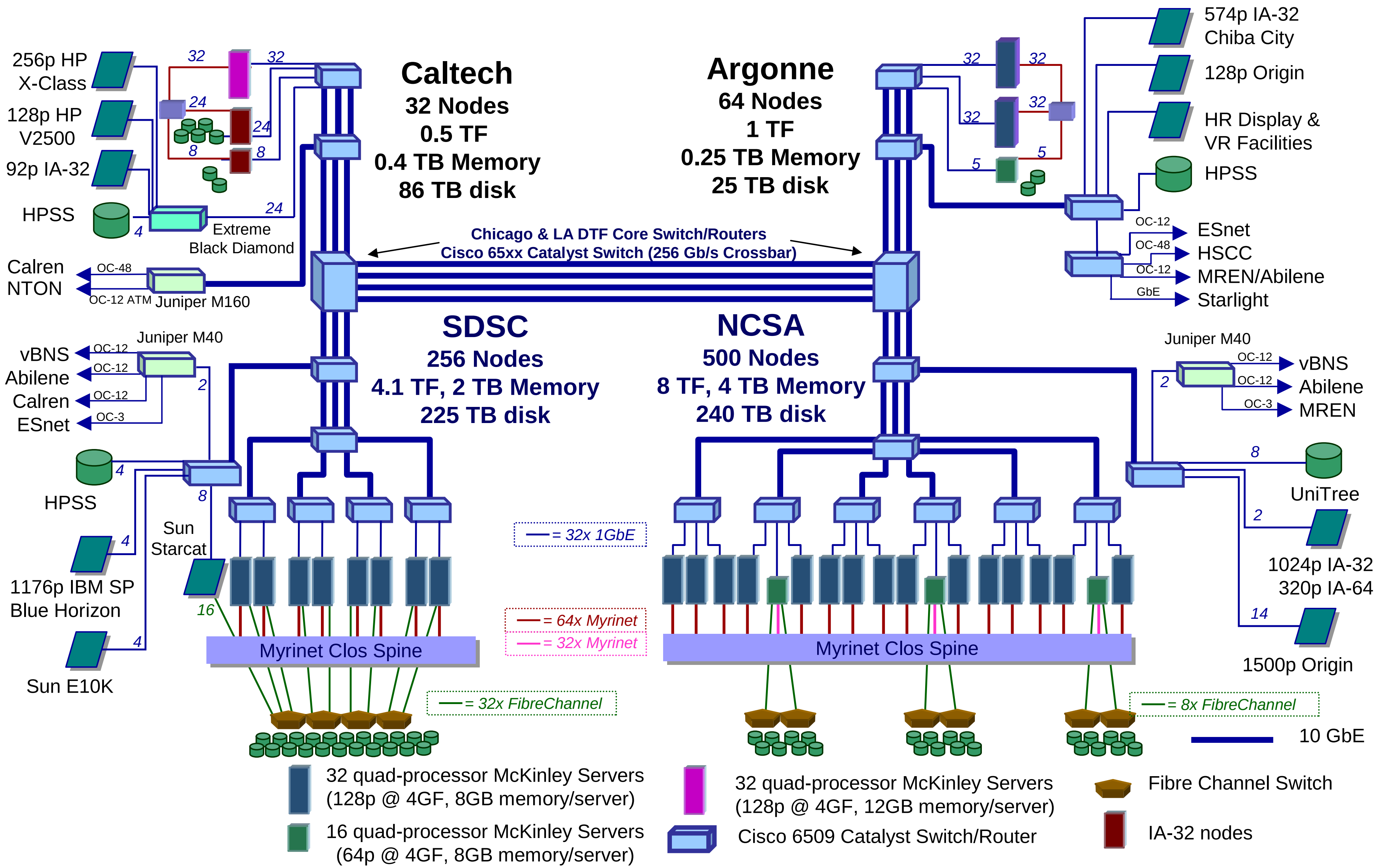
- DTF's project goal is deployment of a production Grid environment
  - staged deployment based on service priorities
  - first priority is a linked set of working IA-64 based clusters
    - immediately useful by the current NSF PACI user base
    - supporting current high-end applications
    - standard cluster and data management software
  - Grid software deployed in phases
    - basic, core, and advanced services

- DTF technology choices based on application community trends
  - > 50% of top 20 PACI users compute on Linux clusters
    - development and production runs
  - majority of NSF MRE projects plan Data Grid environments

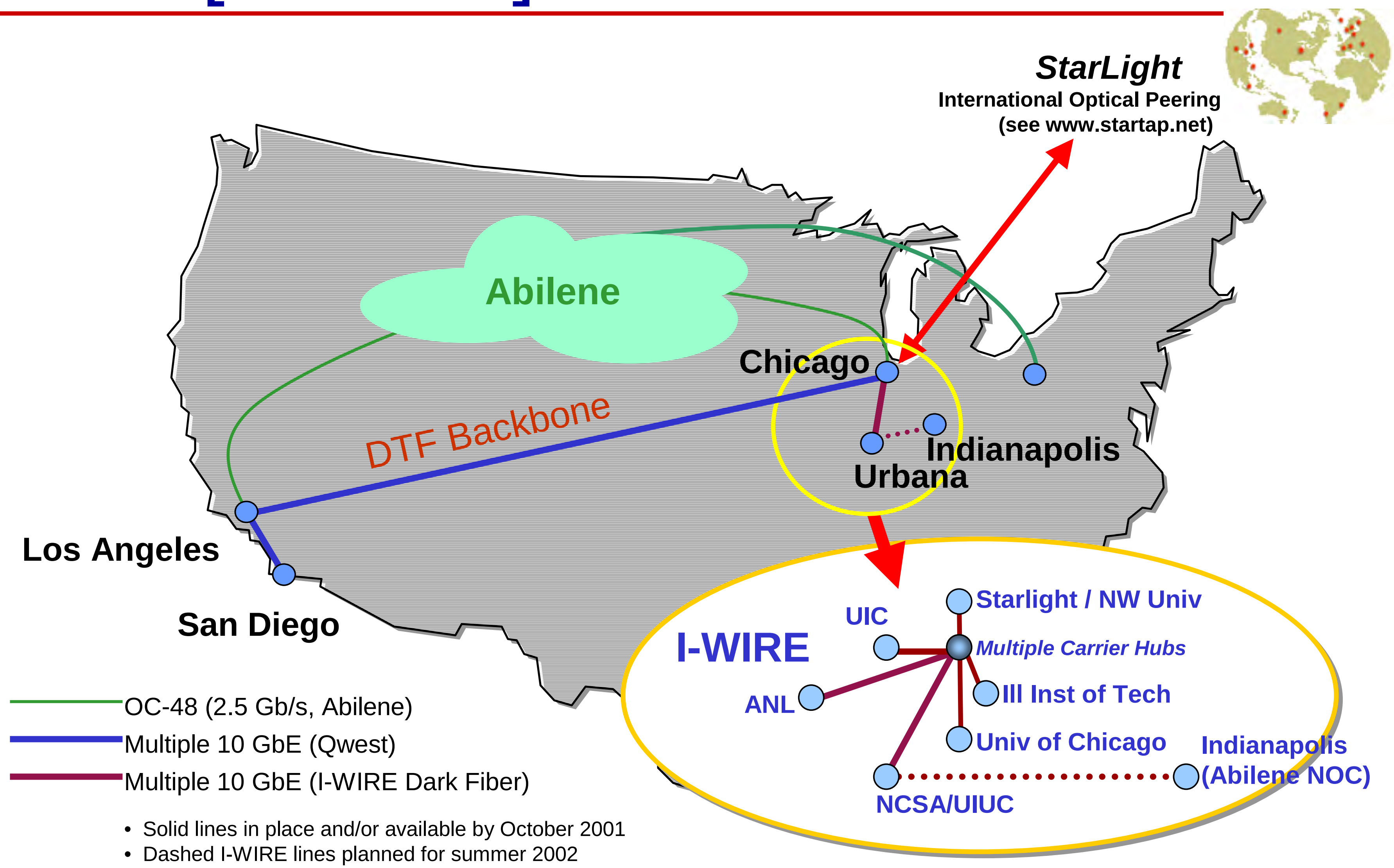
Major DTF and TeraGrid Tasks

- Create management structure – Harder than we thought!
- Engage major application teams – Starting with ITRs and MREs
- Construct high bandwidth national network – On track
- Integrate terascale hardware and software – Planning underway
- Establish distributed TeraGrid operations – New Concepts needed
- Deploy and harden Grid software – Need Grid testbeds
- Expand visualization resources – Development needed
- Implement outreach and training program – PACI Leverage
- Assess scientific impact – Need metrics and process

NSF PACI 13.6 TF Linux TeraGrid



TeraGrid [40 Gbit/s] DWDM Wide Area Network



- Solid lines in place and/or available by October 2001
- Dashed I-WIRE lines planned for summer 2002

TeraGrid Middleware Definition Levels

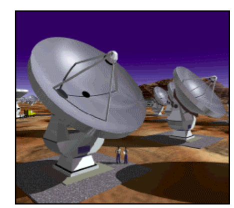
- Basic Grid services [little new capability]
  - deployment ready
  - in current use
  - immediate deployment planned
- Core Grid services [essential Grid]
  - largely ready
  - selected hardening and enhancement
  - planned deployment in year one
- Advanced Grid services [True Grid]
  - ongoing development
  - expect to deploy in year two and later

Grid Middleware [toolkits for building Grids]

- PKI Based Security Infrastructure
- Distributed Directory Services
- Reservations Services
- Meta-Scheduling and Co-Scheduling
- Quality of Service Interfaces
- Grid Policy and Brokering Services
- Common I/O and Data Transport Services
- Meta-Accounting and Allocation Services

Expected NSF TeraGrid Scientific Impact

- Multiple classes of user support
  - each with differing implementation complexity
    - minimal change from current practice
    - new models, software, and applications
- Benefit to three user communities
  - existing supercomputer users
    - new capability [FLOPS, memory, and storage]
  - data-intensive and remote instrument users
    - linked archives, instruments, visualization and computation
    - several communities already embracing this approach
      - GriPhyN, BIRN, Sloan DSS/NVO, BIMA, ...
  - future users of MRE and similar facilities
    - DTF is a prototype for ALMA, NEESGrid, LIGO, and others



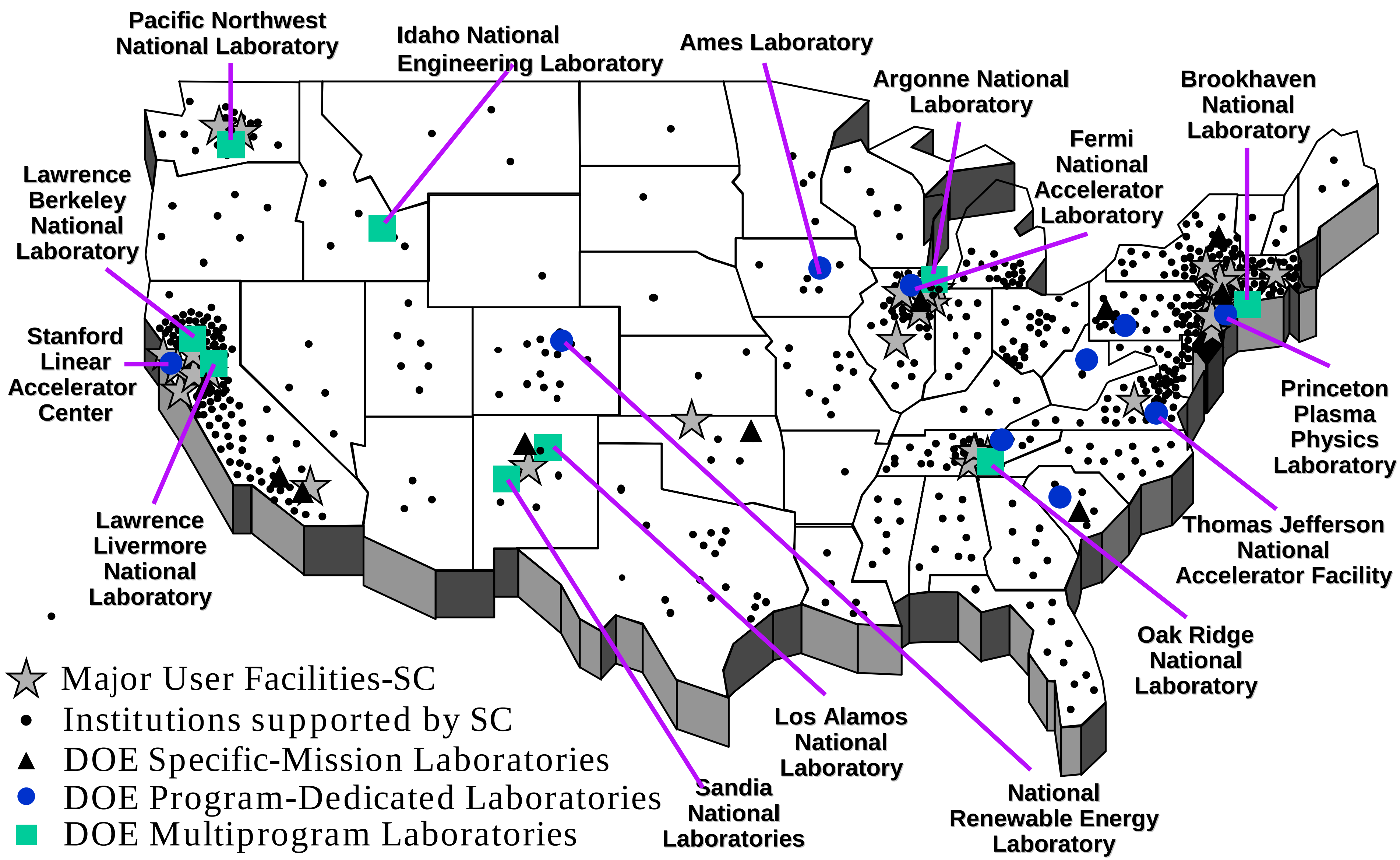
NEON
National Ecological Observatory Network

Strategies for Building Computational Grids

- Three current approaches to developing and deploying Grid Infrastructure
  - Top Down – NASA IPG, NSF DTF, UK e-science
  - Bottom Up – Life Science's Web Based Computing
  - User Community Based – GriPhyN, iVDG, PPDG, etc.
- Current Grid Software R+D Mostly focused on Top Down and User Community Models
- Major Grid Building "activities"
  - Grid software infrastructure and toolkit development
  - Grid hardware resources [systems, networks, data, instruments]
  - Grid applications development and deployment
  - Grid resource allocation and policy development

DOE Programs and Facilities

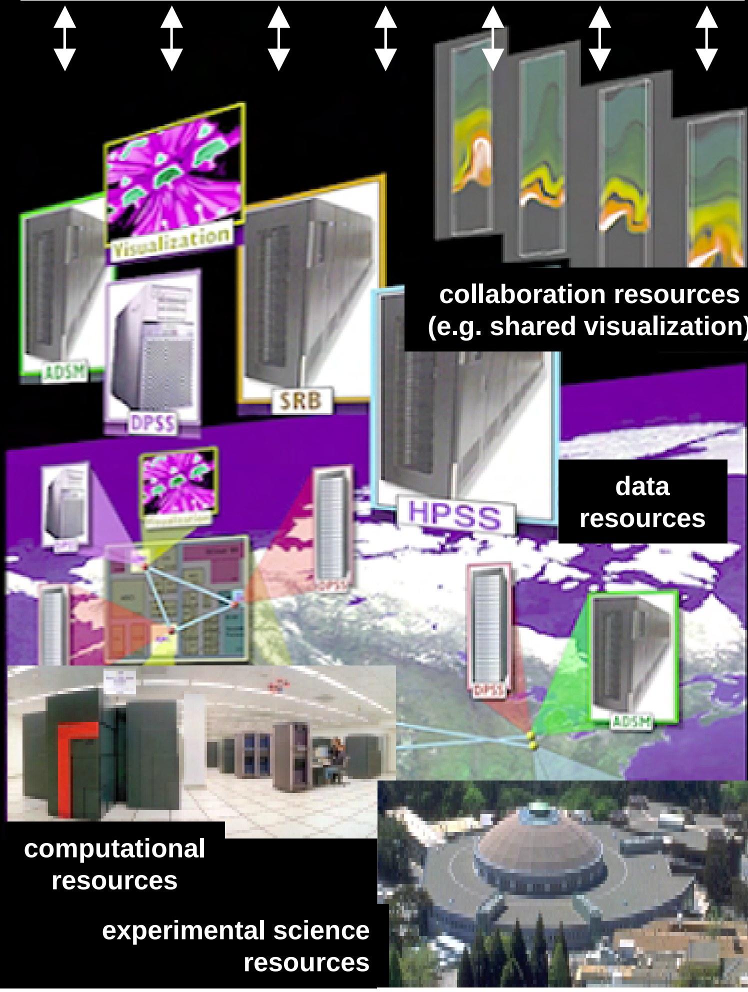
Technical Challenges: Distributed Resources, Distributed Expertise



Vision for a DOE Science Grid

Scientific applications use workflow frameworks to coordinate resources and solve complex, multi-disciplinary problems

Grid services provide uniform access to many diverse resources



Large-scale science and engineering is typically done through the interaction of

- collaborators
- heterogeneous computing resources,
- multiple information systems, and
- experimental science and engineering facilities

all of which are geographically and organizationally dispersed.

The overall motivation for "Grids" is to enable routine interactions of networked combinations of these resources to facilitate large-scale science and engineering.

high-speed network services and infrastructure provide the substrate for Grids

Two Primary Goals

- Build a DOE Science Grid that ultimately incorporates computing, data, and instrument resources at most, if not all, of the DOE Labs and their partners.

- Advance the state-of-the-art in high performance, widely distributed computing so that the Grid can be used as a single, very large scale computing, data handling, and collaboration facility.

Grid Strategies Appropriate for DOE-SC

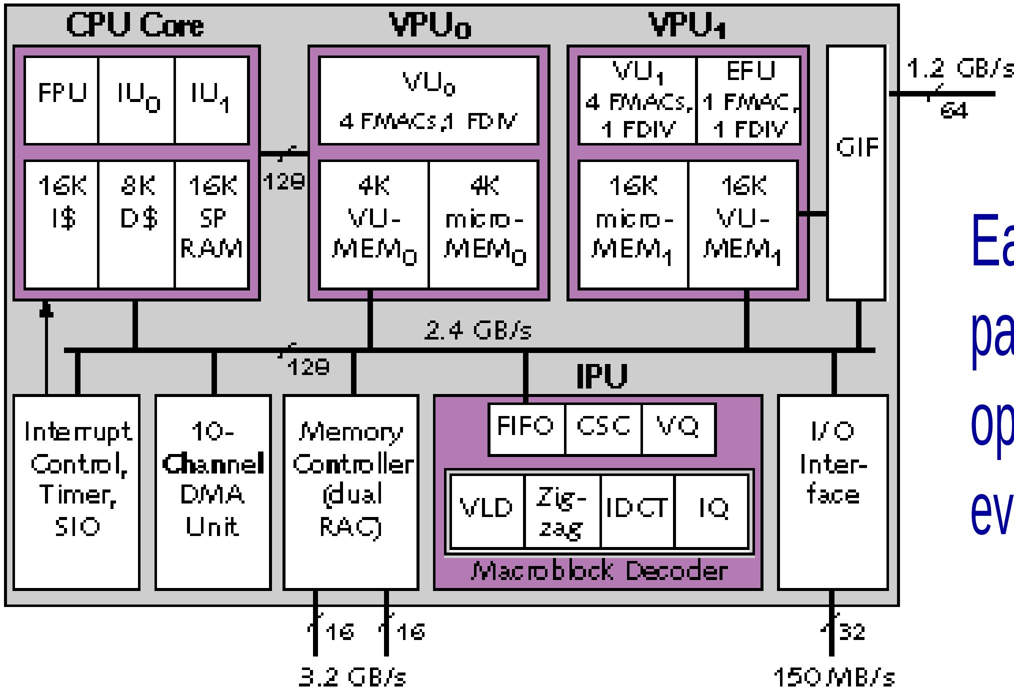
- Possible three layered structure of DOE-SC Grid resources
  - Large-Scale Grid Power Plants O[10]
    - [Multiprogram Labs, LBNL-NERSC, ORNL, ANL, PNNL, etc.]
  - Data and Instrument Interface Servers O[100]
    - [Major DOE Facilities, LHC, APS, ALS, RHIC, SLAC, etc.]
  - PI and small "l" laboratory based resources O[1000-10,000]
    - [Workstations and small Clusters, laboratory data systems, databases, etc.]
- Need <u>tool development, applications development and support</u> appropriate to each layer and user community
- Need <u>resource allocation policies</u> appropriate for each class of resource and user community

Near Term Directions for "Clusters"

- High-Density Web Server Farms [IA-32, AMD, Transmeta]
  - Blade based servers optimized for dense web serving
  - Scalable, but not aimed at high-performance numerical computing
- Passive Backplane Based Clusters [IA-32, Infiniband]
  - Reasonably dense packaging possible
  - High-Scalability not a design goal
- IA-64, x86-64 and Power4 "Server" based compute nodes
  - Good price performance, poor packaging density
  - Designed for commercial I/O intensive configurations
- Sony Playstation2 [Emotion Engine, IBM Cell Project]
  - Excellent pure price performance $50K/Teraflop
  - Not a balanced system, difficult microarchitecture

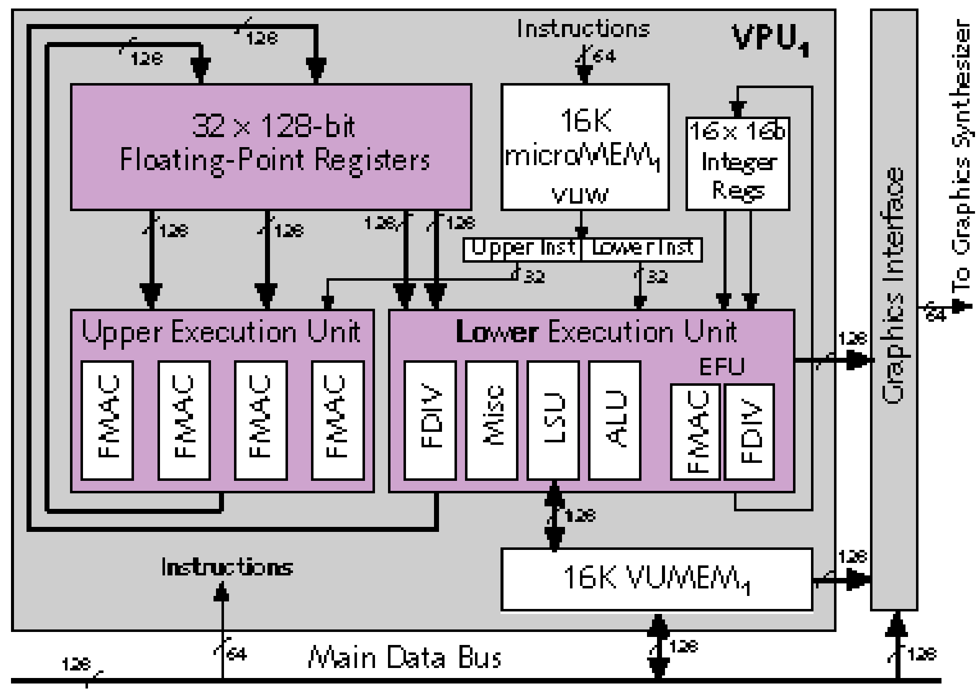
Limits to Cluster Based Systems for HPC

- Memory Bandwidth
  - Commodity memory interfaces [SDRAM, RDRAM, DDRAM]
  - Separation of memory and CPU implementations limits performance
- Communications fabric/CPU/Memory Integration
  - Current networks are attached via I/O devices
  - Limits bandwidth and latency and communication semantics
- Node and system packaging density
  - Commodity components and cooling technologies limit densities
  - Blade based servers moving in right direction but are not High Performance
- Ad Hoc Large-scale Systems Architecture
  - Little functionality for RAS
  - Commodity design points don't scale



Each vector unit has enough parallelism to complete a vertex operation [19 mul-adds + 1 divide] every seven cycles.

The PSX2's Emotion Engine provides ten floating-point multiplier-accumulators, four floating-point dividers, and an MPEG-2 decoder.



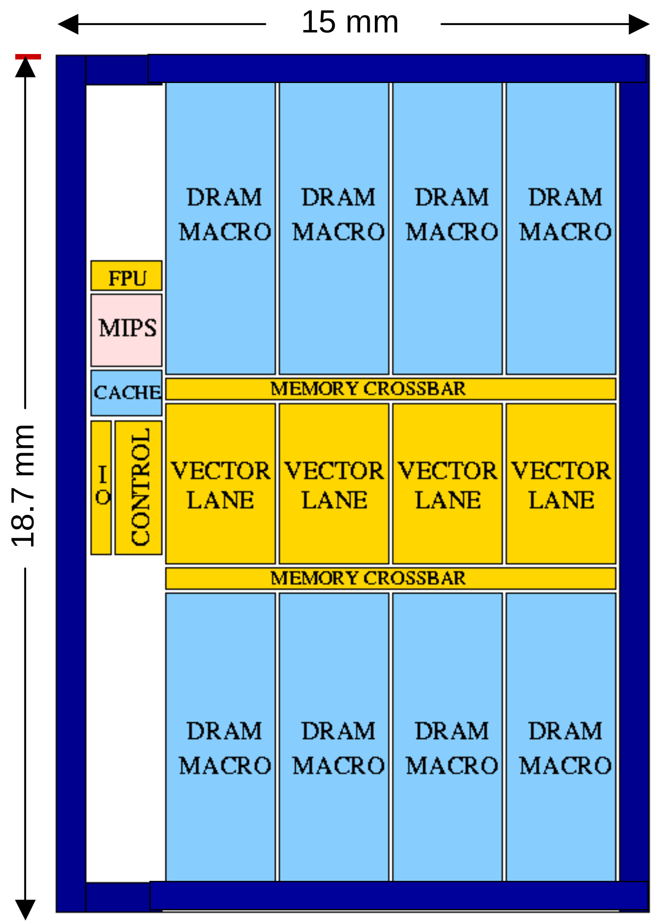
IBM, Sony and Toshiba "Cell" Project

- $400M investment towards Teraflop processor
- Targeted at PS3, broadband applications
  - Each company will produce products based on the core technology
- 100 nm feature size ⇒ 2006 [based on SIA Roadmap]
- Design Center in Austin TX opening later this year
- Sony's description of PS3 is <u>1000x performance of PS2</u>
  - Planned use for all of Sony's product lines
  - Video, Audio, Computer Games, PCs Etc.

Cluster Technology Will not Scale to Petaflops

- Affordable and Usable Petaflops will require improvements in a number of areas:
  - Improved CPU-Memory Bandwidth [e.g. PIM, IRAM]
  - HW Based Latency Management [e.g. multithreaded architectures]
  - Integrated Communications Infrastructure [e.g. on-chip networking]
  - Increased level of system Integration and Packaging [e.g. SOC, CIOC]
  - New Large-scale Systems Architectures
    - Aggressive Fault Management and Reliability Features
    - Scalable Systems Management and Serviceability Features
  - Dramatic Improvements in Scalable Systems Software

UCB VIRAM-1 Integrated Processor/Memory



Thanks to DARPA: funding
IBM: donate masks, fab
Avanti: donate CAD tools
MIPS: donate MIPS core
Cray: Compilers, MIT:FPU

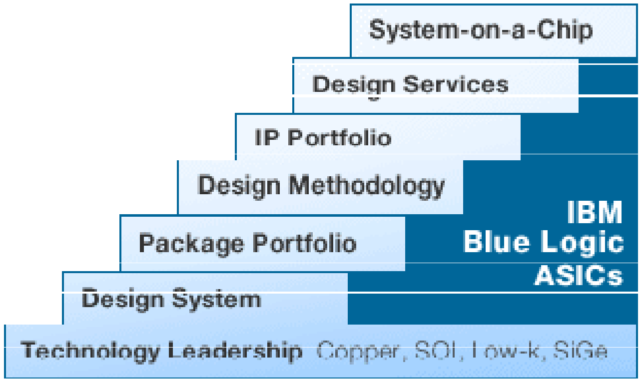
- Microprocessor
  - 256-bit media processor [vector]
  - 14 MBytes DRAM
  - 2.5-3.2 billion operations per second
  - 2W at 170-200 MHz
  - Industrial strength compiler
- 280 mm^2 die area
  - 18.72 x 15 mm
  - ~200 mm^2 for memory/logic
  - DRAM: ~140 mm^2
  - Vector lanes: ~50 mm^2
- Technology: IBM SA-27E
  - 0.18µm CMOS
  - 6 metal layers [copper]
- Transistor count: >100M
- Implemented by 6 Berkeley graduate students

Petaflops from V-IRAM?

- In 2005 .. V-IRAM 6 Gflops/64 MBs
- 50 TF per 19" Rack [~10K CPUs/Disk assemblies per rack]
  - 100 drawers of 100 processors [like library cards]
  - cross bar interconnected at Nx 100 MB/s
- 20 optically interconnected Racks
  - 10^{15} FLOPS $\Rightarrow$~\$20M
- Power ~ 20K Watts x 20 = 400K Watts

A Possible Path Towards Petaflops User Facilities

- Community Based Approach to Petaflops Systems Development
  - Laboratories, Universities, and Applications Communities
  - User Requirements, Software and Systems Design
- Exploiting New Design Ideas and Technology for Scalability
  - Cluster-on-a-chip Level Integration
  - Hardware/software Co-design
- Affordable Petaflops Enable Personal Teraflops
  - $50M PFlops System ⇒ $50K TFlops
  - Enable Broad Deployment and Scientific Impact
- Advanced Networking and Middleware
  - Embed Petaflops Capability in the Grid



Grid Implications of Affordable Petaflops

- \$50M Petaflops system ⇒ \$50K Teraflops systems
- DOE Computing Facilities Circa 2006-2010
  - Power Plant Level ⇒ O[1-10] PFs computers
  - Data Server Level ⇒ O[20-100] 100 TFs Data Servers
  - PI Level ⇒ O[1000-10,000] 1 TF lab systems
- Data Server Capability
  - Power Plant Level ⇒ 10-20 PB secondary, 100-1000 PB tertiary
  - Data Server Level ⇒ 100-500 PB sec, 1000+ PB tertiary
  - PI Level ⇒ ~1 PB secondary
- Networking Capability
  - Ideal BW ⇒ 10% of bisection bandwidth per system for peer-to-peer
  - Terabit WAN backbones and backplanes will be needed

Summary

- Grid Based Computing Concept is well matched to DOE’s Distributed Facilities and Missions needs
- Grids do not replace need for large-scale computers
  - Increases high-end demand via Portals
  - Increases data intensive computing and high-performance networking
  - software environments link desktops to high-end platforms [petaflops]
- Grids require *new* ways to allocate and manage computing and data resources
  - Need a broader view of resources and resource allocations
- Grids and Technology for Petaflops Facilities make sense together
  - Technologies for Petaflops will power future grids at all levels

Recommendations

- DOE should aggressively pursue development of Grid Technologies and deployment of Grid based Infrastructure
- DOE should facilitate Grid Applications Communities relevant to mission areas: Security, Energy, Climate, HEP, etc.
- DOE should participate in National and International coordination of Grid development and Deployment
- DOE should support development of computing platform technologies that will enable future Grid engines, including new approaches to Petaflops and associated affordable Teraflops