Shared Computing Resource for Advancing Science and Engineering Using the Cray XD1 at Rice University

Jan E. Odegard

Executive Director

Computer & Information Technology Institute

Rice University

odegard@rice.edu; +1.713.348.3128

http://www.citi.rice.edu

Cray User Group Technical Conference – Lugano, Switzerland

May 8-11, 2006

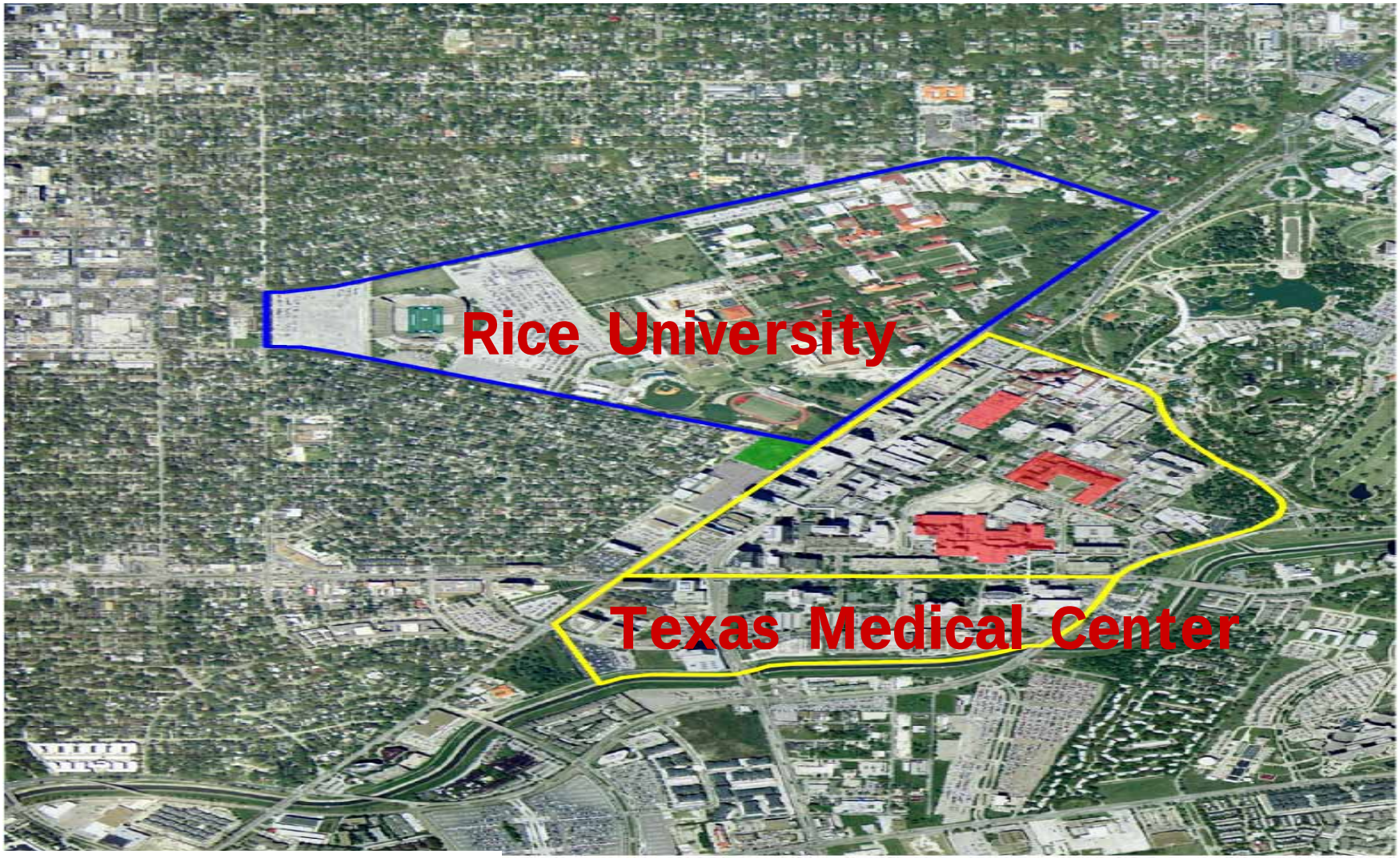
Computer and Information Technology Institute

RICE University

Rice University

- Independent, coeducational, private
- ~540 faculty, 2800 undergraduates, 1900 graduate students (5:1 undergrad student/faculty ratio)
- $3.6 billion endowment
- Rice undergraduates – National Merit Scholars (highest % of all national universities & colleges last ten years!)
- Best educational value – consistently ranked 1st or 2nd
- Consistently rated in top 20 Universities by USNews
- 10 faculty – National Academy of Sciences/Medicine
- 14 faculty – National Academy of Engineering

Where is Rice University?



RICE University

Computer and Information Technology

To build a community of scholars that engages in collaborative research and education covering virtually every aspect of information technology and computing

Directors:

Ken Kennedy (1986-1992)

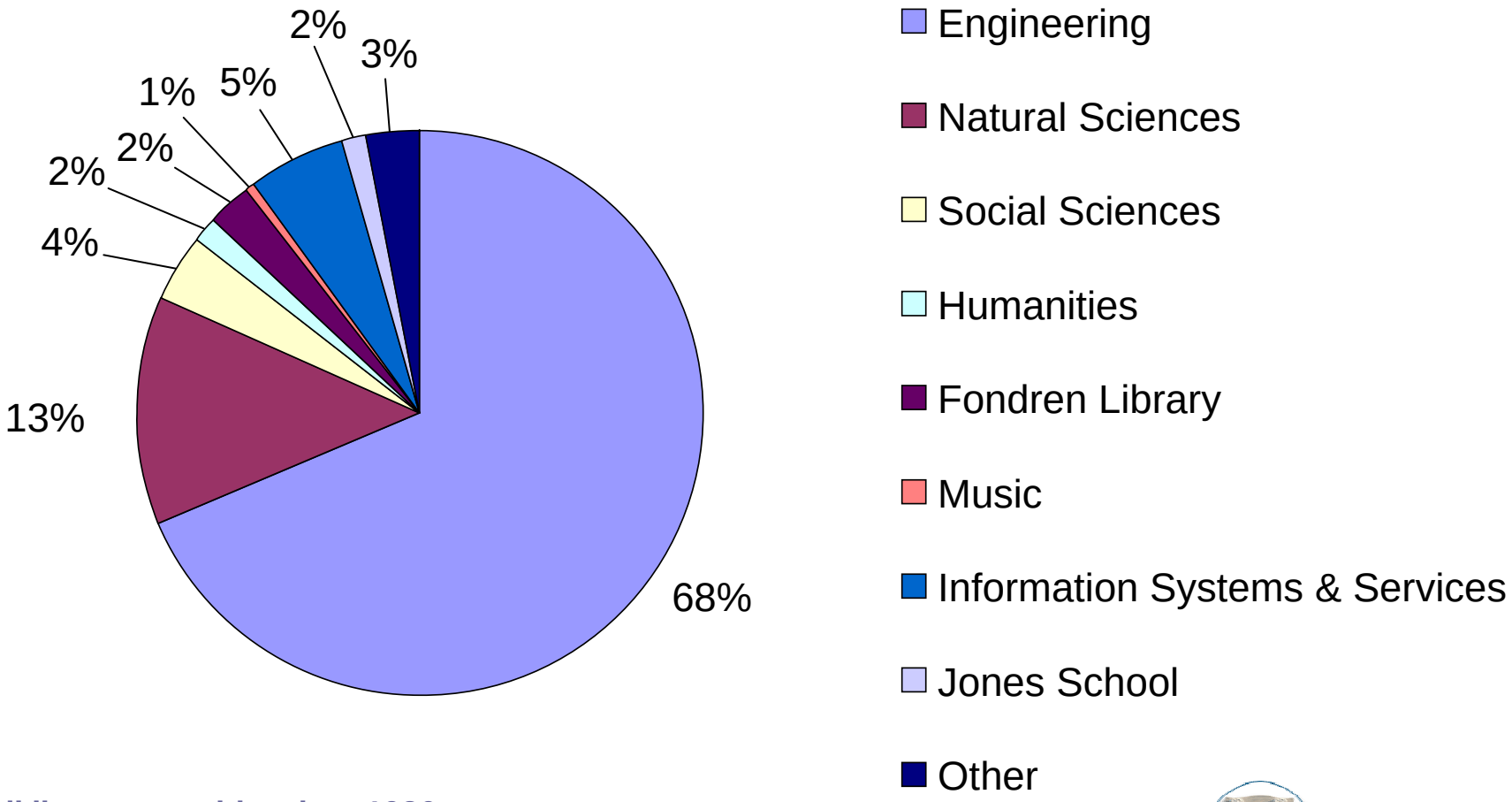
Sidney Burrus (1992-1998)

Willy Zwaenepoel (1998-2001)

Moshe Vardi (2001-…)

CITI Members

6 schools ←→ 20 departments ←→ ~130 members
7 centers ←→ ~15 ad hoc research groups

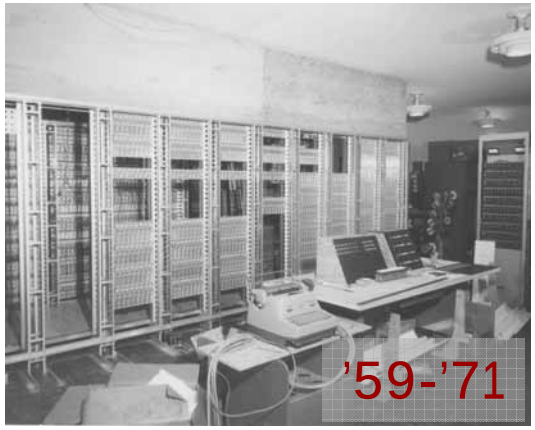


Research Centers

- Center for High Performance Software (HiPerSoft)
  - Director: Ken Kennedy, CS
- Center for Multimedia Communication (CMC)
  - Director: Behnaam Aazhang, ECE
- Center for Computational Geophysics (CCG)
  - Co-directors: B. Symes, CAAM / A. Levander, ES
- Center for Computational Finance & Economic Systems (CoFES)
  - Director: Kathy Ensor, STAT
- LAboratory for NanoPhotonics (LANP)
  - Director: Naomi Halas, ECE
- Center for Technology in Teaching and Learning (CTTL)
  - Director: Tony Gorry, CS
- Center for Excellence and Equity in Education (CEEE)
  - Director: Richard Tapia, CAAM

Shared Research Computing ...

- From the archives:

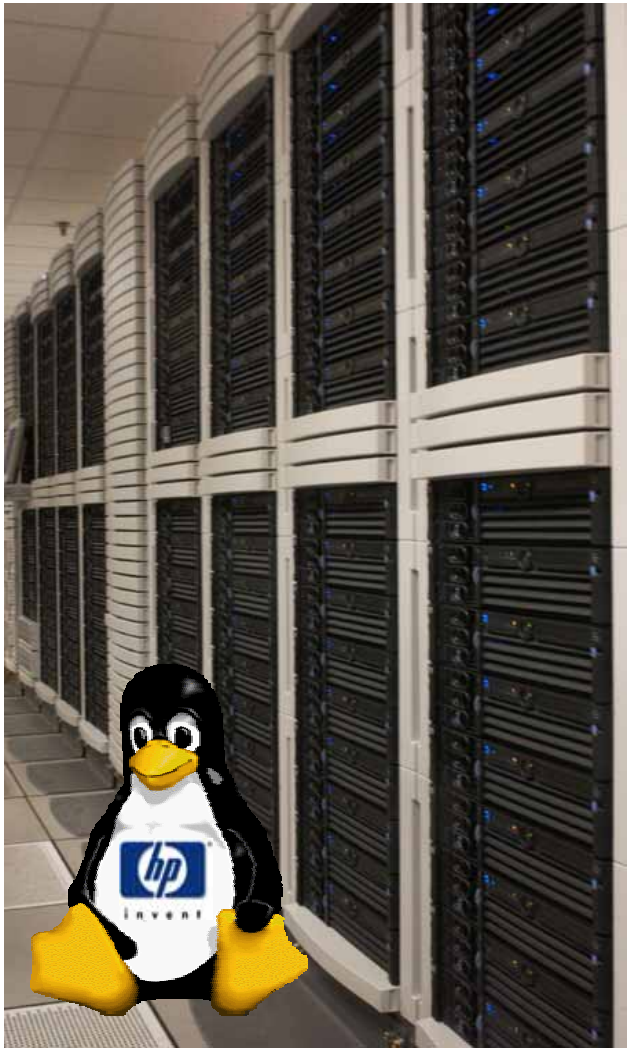


- and many other systems from past and current…
- January 2002:
  - CITI pulled together a team of ~30 investigators and wrote a successful NSF MRI proposal
- January 2004:
  - CITI pulled together another team of ~35 investigators and wrote a second successful NSF MRI proposal

HP Integrity
Intel Itanium

"Rice Terascale Cluster"

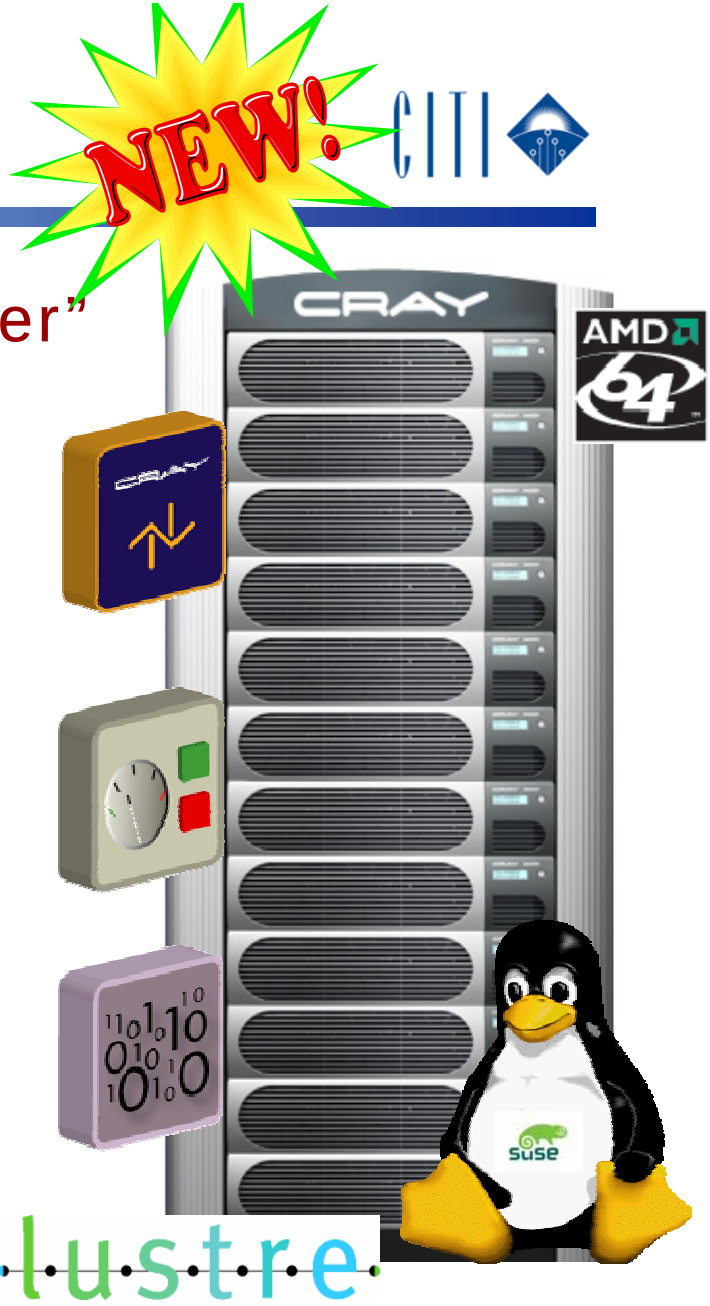
- ~1 TeraFLOP HP Linux cluster*
  - 286 Intel® Itanium® 2 processors
    - 900MHz, 1.5MB
    - HP zx1 chipset
    - 134 dual nodes
    - 5 quad nodes
  - Myrinet 2000 (96 nodes)
  - Foundry Network – GigE (all)
  - 640GB memory
  - 11TB Disk
    - 6.5TB on node
    - 1TB scratch back-end
    - 3.5TB shared front-end



NSF MRI, Rice, Intel and HP

Cray XD1 System

Dual-Core AMD Opteron™



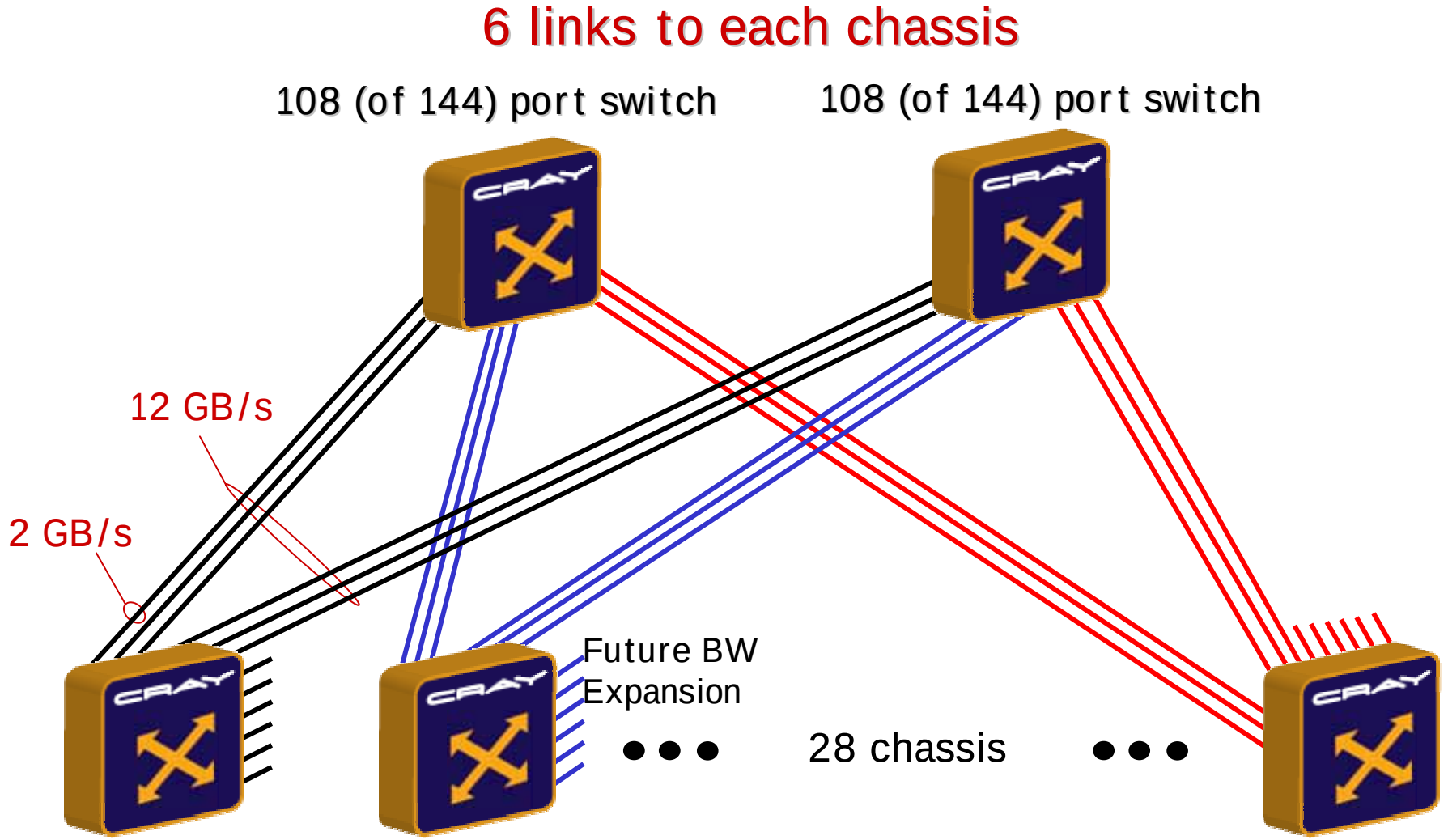
"Rice Computational Research Cluster"

- ~3 TeraFLOP Cray XD1 Linux cluster*
  - 336 Dual-Core AMD Opteron™ 275
    - 2.2GHz, 1MB / Core
    - 168 dual socket nodes (4 way SMP)
    - 8 GB DDR 400 / compute SMP
    - 16 GB DDR 333 / system SMP
  - Cray RapidArray (4x Infiniband)
  - 1.4 TB DDR2 400
  - 12 TB Local Disk
  - 6 TB Lustre parallel file system
  - 10 TB NFS file system
  - One XD1 Chassis with FPGA
    - 6 Xilinx Virtex-4/LX160

NSF MRI, Rice, AMD and Cray

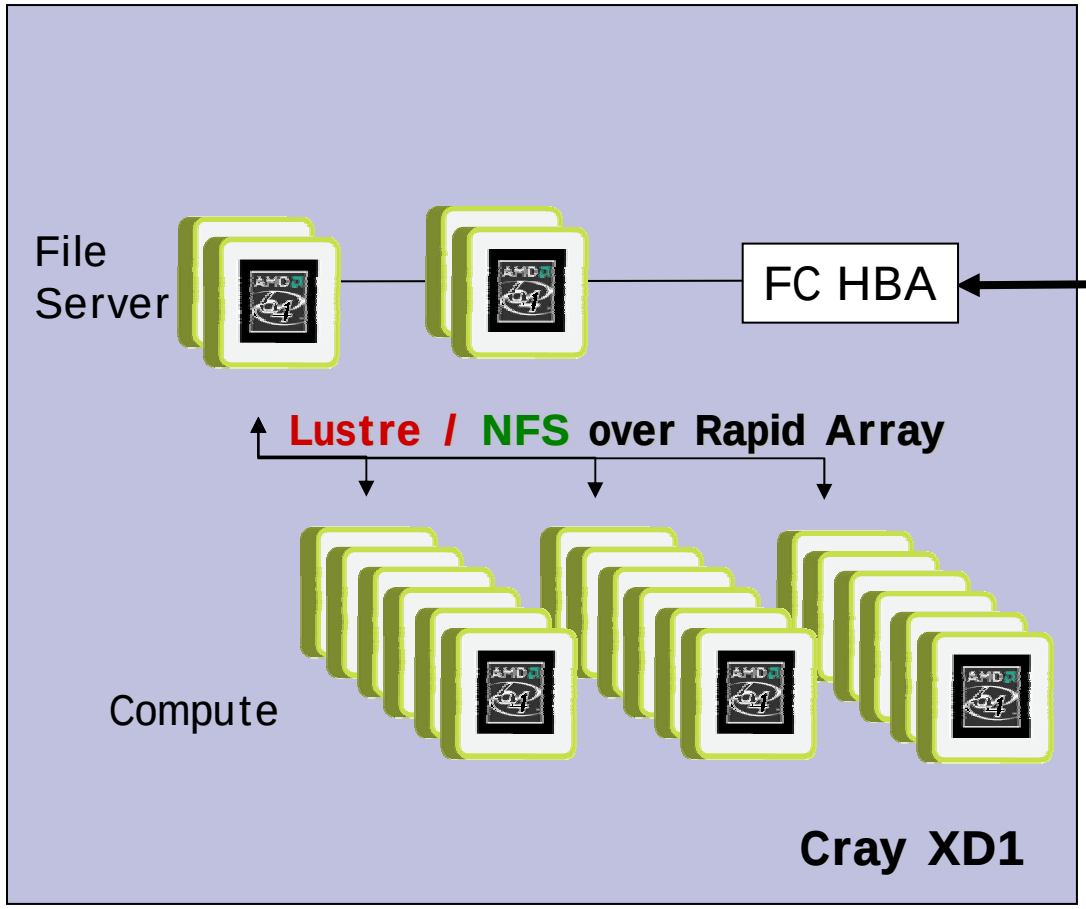
Rice's Cray XD1

Rice's Fat Tree Topology



File Systems: External Storage

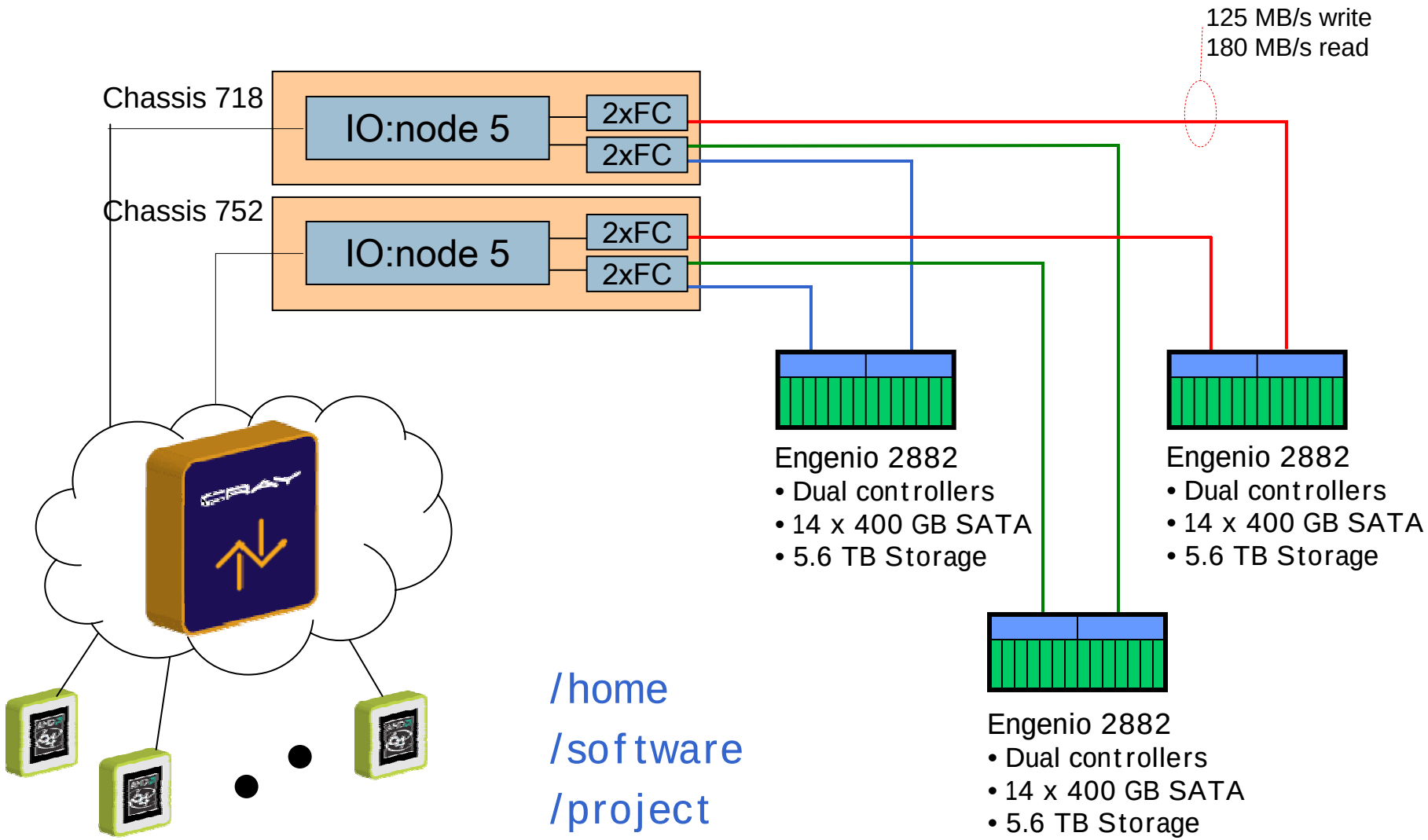
SMP acting as a File Server



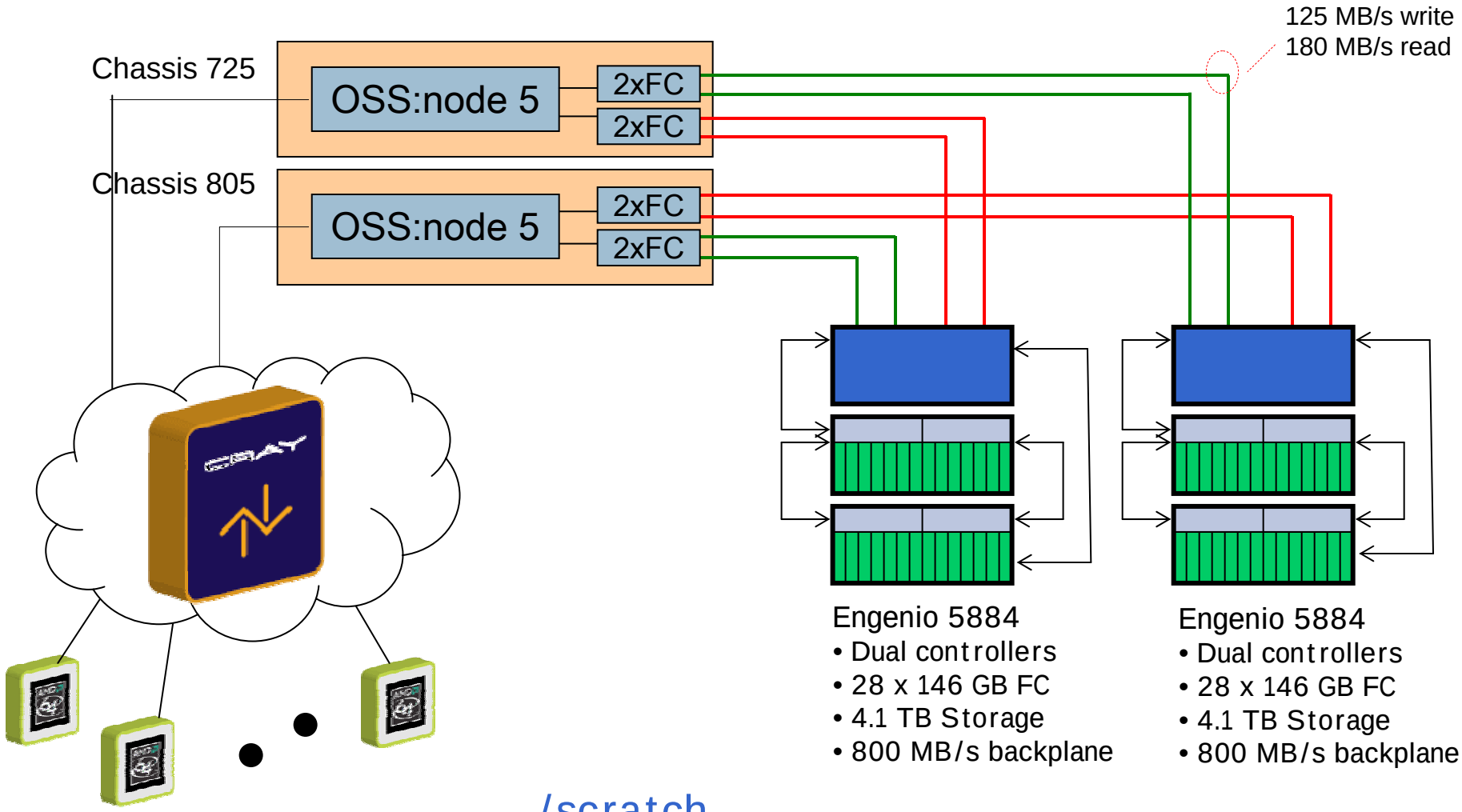
Engenio Storage Subsystems With SATA / FC Disks

Global File Systems:
Lustre File System
NFS File System

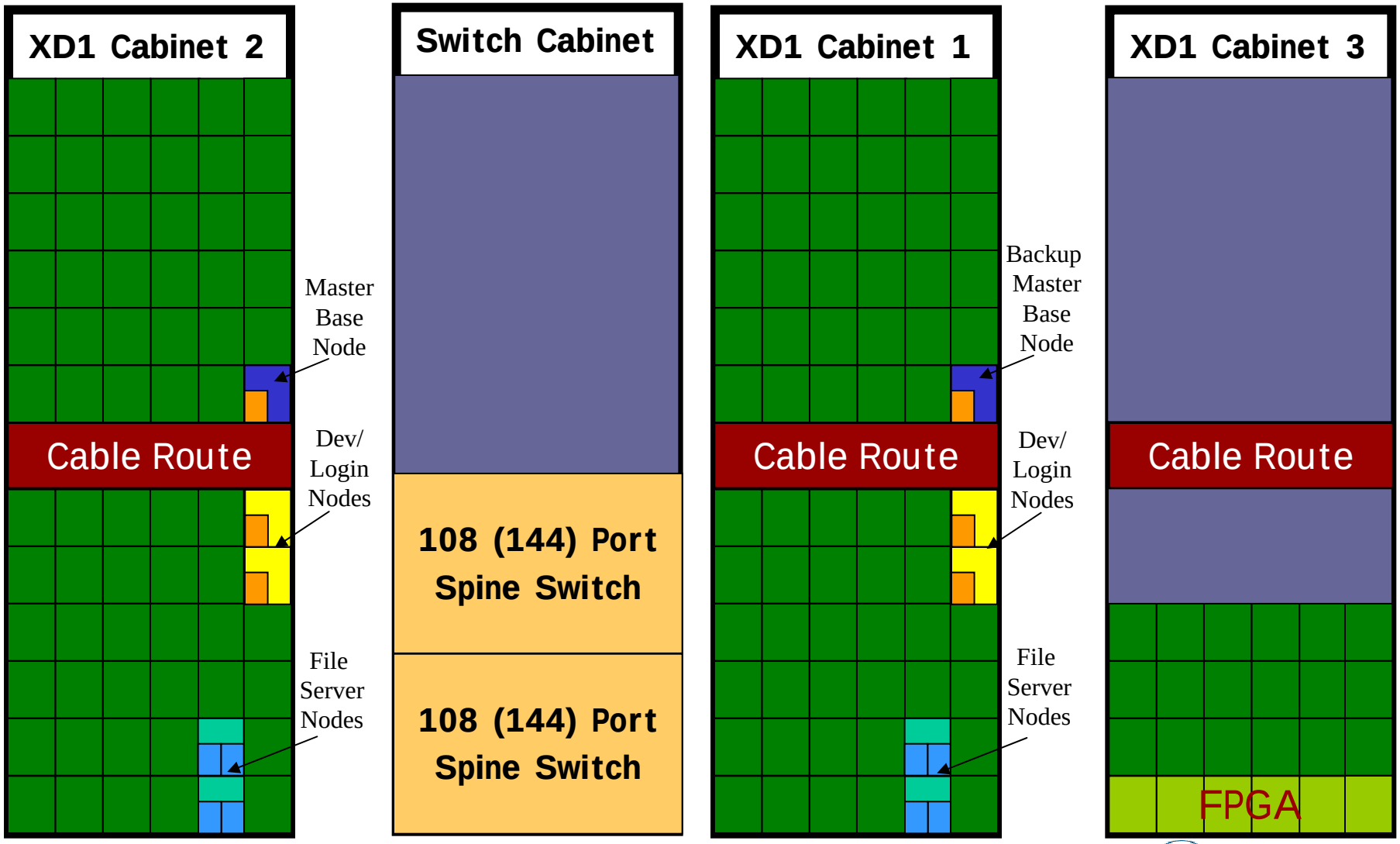
NFS Configuration



Lustre Configuration



Cabinet & Chassis Layout

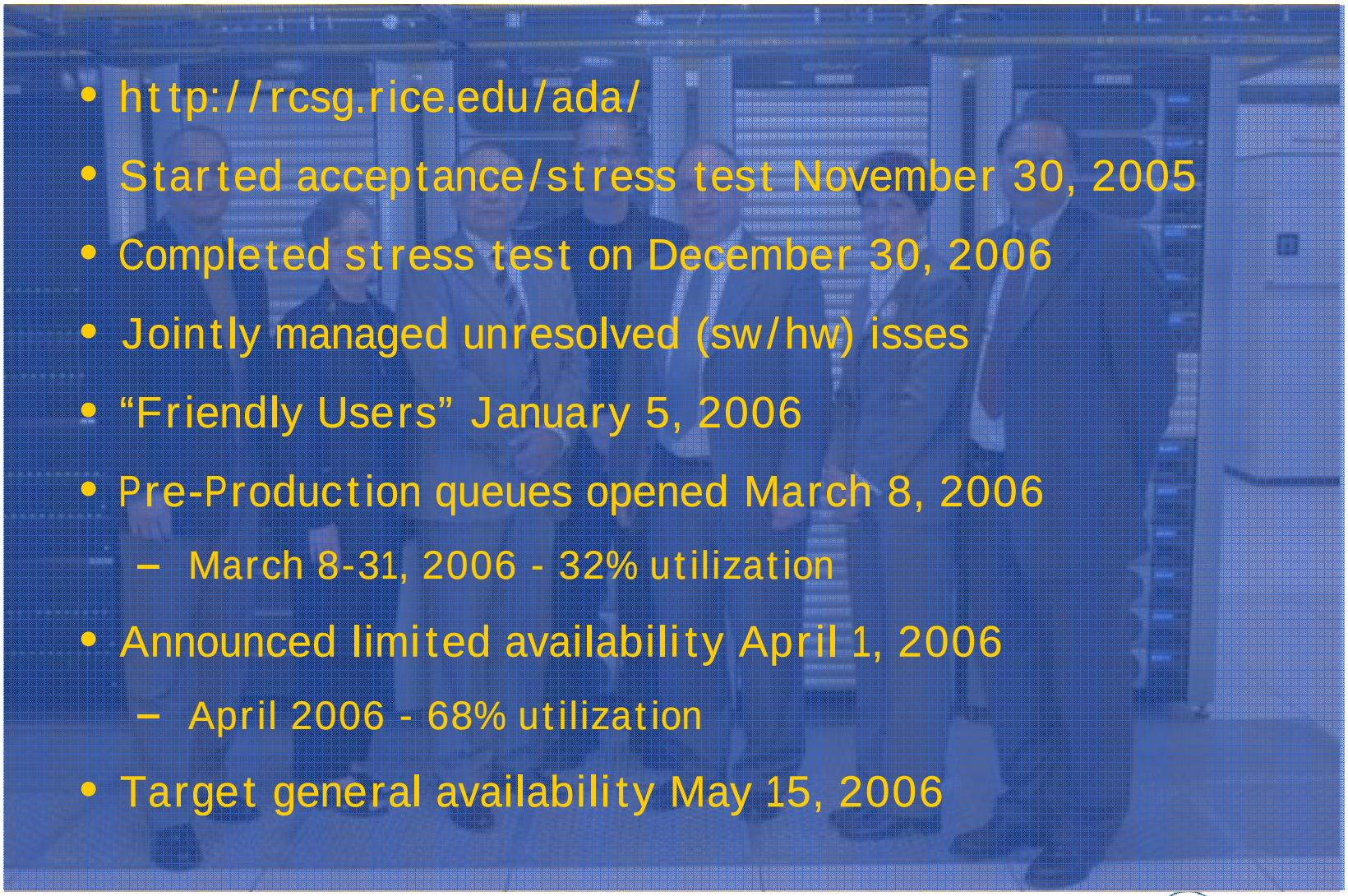


Rear View

Acceptance Criteria

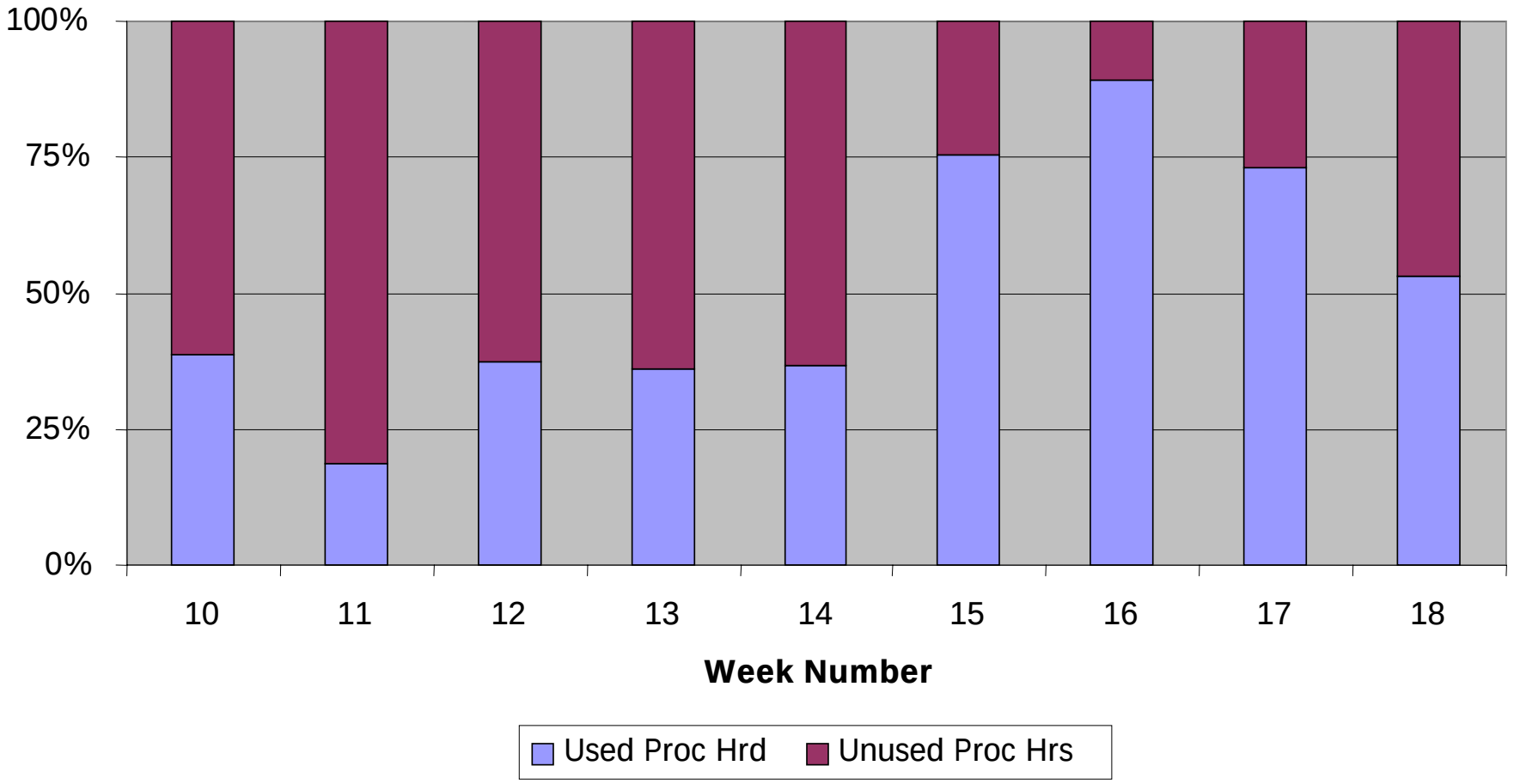
- Step 1: Pre Ship Test
  - Run a scaled down version of the acceptance test before system could ship
- Step 2: Site Install Test
  - Installation validation test
  - Show performance >= performance achieved before shipping
- Step 3: Production Test (29-30-60)
  - Cray responsible for system availability
  - Rice responsible for production work-load
  - Deliver production capabilities 29 of 30 consecutive days within a 60 day window

February 10, 2006: Ribbon Cutting

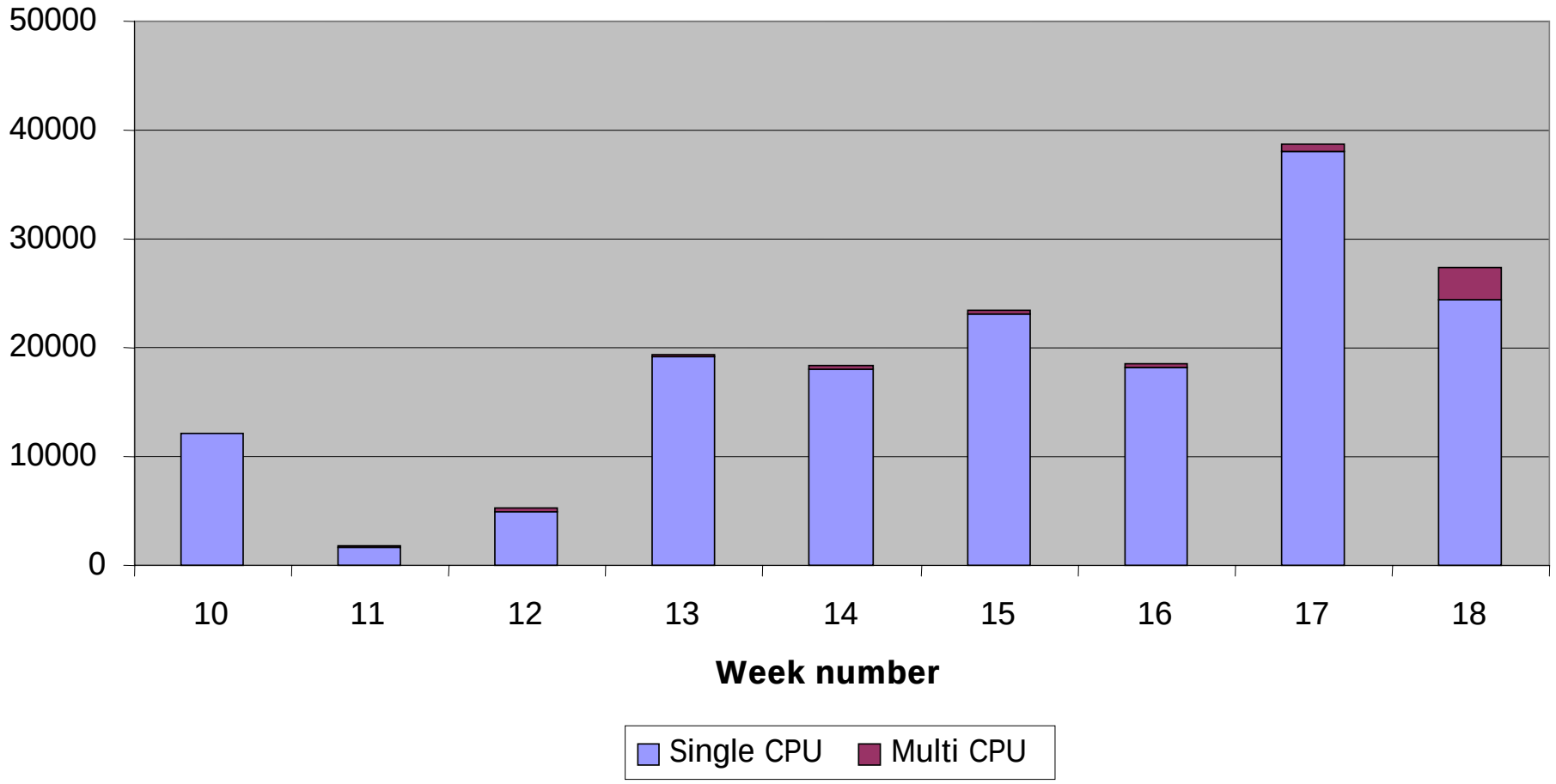


- http://rcsg.rice.edu/ada/
- Started acceptance/stress test November 30, 2005
- Completed stress test on December 30, 2006
- Jointly managed unresolved (sw/hw) isses
- “Friendly Users” January 5, 2006
- Pre-Production queues opened March 8, 2006
  - March 8-31, 2006 - 32% utilization
- Announced limited availability April 1, 2006
  - April 2006 - 68% utilization
- Target general availability May 15, 2006

System Utilization



Getting Research Done!



So who runs the system?

Early or Ongoing Issues

- Power Harmonics
- LDAP integration
- PBS
  - Died at random
    - Likely caused by LDAP instability
  - Code failed
    - File Descriptor hard coded to 1024 → 8192
- IMB did not run consistently on >500 cpus
  - Required more memory buffers than could be stored in RapidArray registration table
  - Triggered bug that caused memory corruption causing subsequent code (small or large) to fail
  - 1.4GA will solve this issue
- LVM at reboot requires manual intervention
  - Requires kernel fix

Early or Ongoing Issues ...

- Unable to move nodes between partitions due to an issue with L2F
  - AM 1.4 supposedly solves this
- Scalability of System Provisioning
  - Unacceptable
- Lustre crashes occasionally
  - Patch being tested at Alabama site
- Base node crashes roughly every 2 weeks
  - Kernel on base logs errors related to AMD IOMMU
  - Patch in the works but it does not support Lustre
- cfengine is a resource hog
  - High load without doing much “useful” work
  - AM 1.4 supposedly addresses this

AM Observations/Recommendations

- AM is a great idea but:
  - AM is by default very intrusive
  - AM would be more acceptable if modularized
    - Not all or nothing
  - AM does not scale well for large installations
  - AM does not provide information comparable to
    - nagios
    - not able to provide historical resource information
    - alarm and fault response not sufficiently flexible

What are users doing with the system?

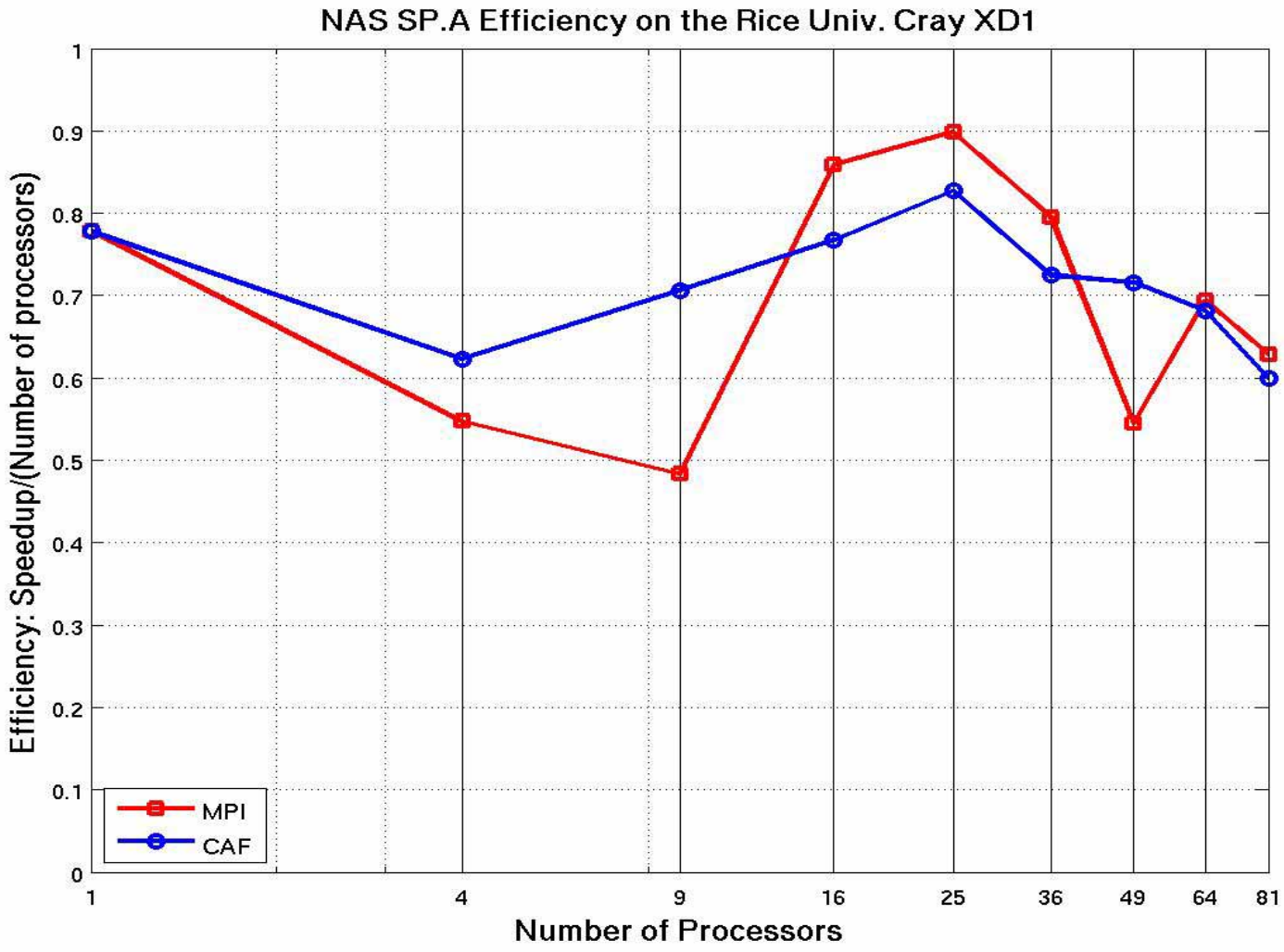
HPCToolkit

- John Mellor-Crummey et al.
- An open-source suite of multi-platform tools for profile-based performance analysis of applications
  - Almost ported
- http://www.hipersoft.rice.edu/hpctoolkit/

- Cristian Coarfa, John Mellor-Crummey et al.
- Porting Co-Array Fortran (CAF) to XD1
  - Port almost completed
  - MPI vs CAF performance results on NAS SP and MG
- http://www.hipersoft.rice.edu/caf/index.html

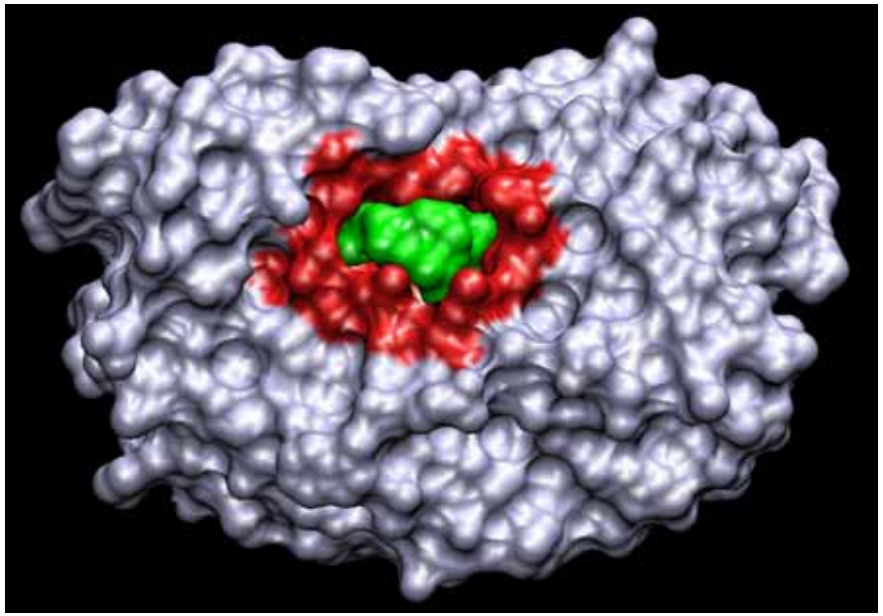
HiPerSoft

NAS SP.A, SP.B, MG.A & MG.B



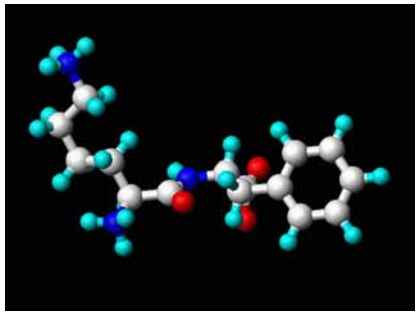
Bioinformatics Research: Computer-Assisted Drug Design

- Lydia Kavraki *et al.*



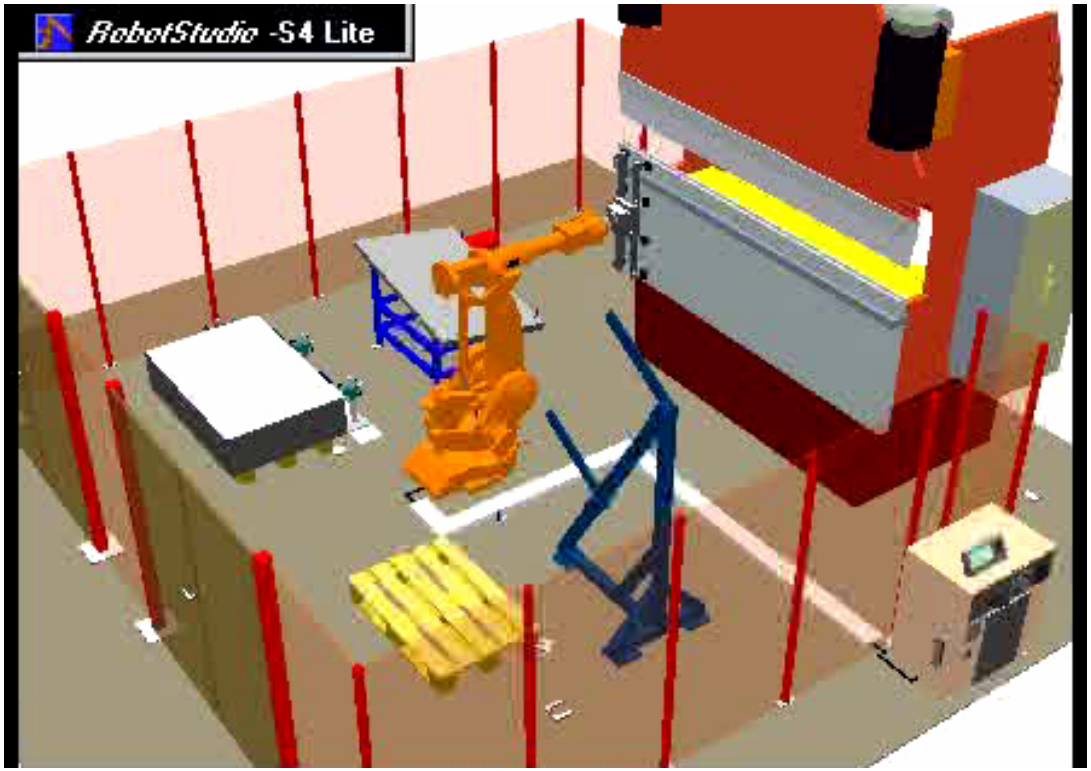
A drug is often obtained by the docking of one molecule (drug) in the "cavity" of a larger molecule (receptor).

Search for likely candidates



GOAL: find a minimum energy best fit

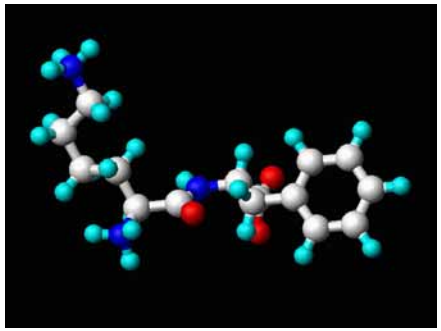
Computer-Assisted Path Planning

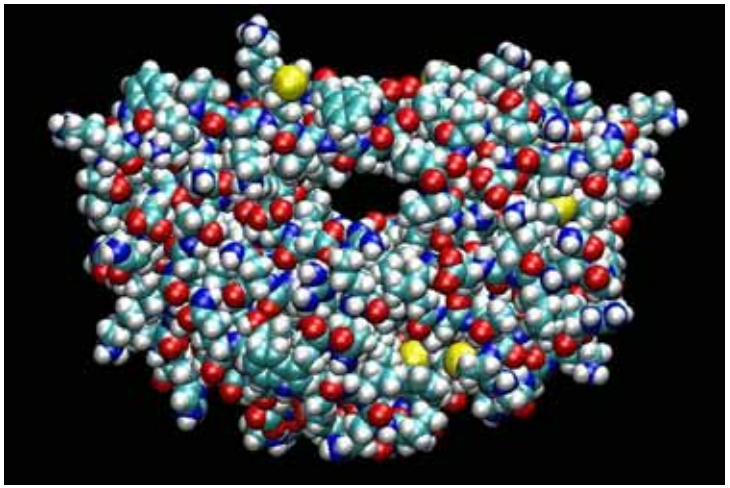


- Planning for a robot moving among obstacles resembles
- Planning for a drug moving in the energy field of a receptor
- Only planning for a drug is much harder…..

A Robotics-inspired Search Approach

Both molecules are flexible!





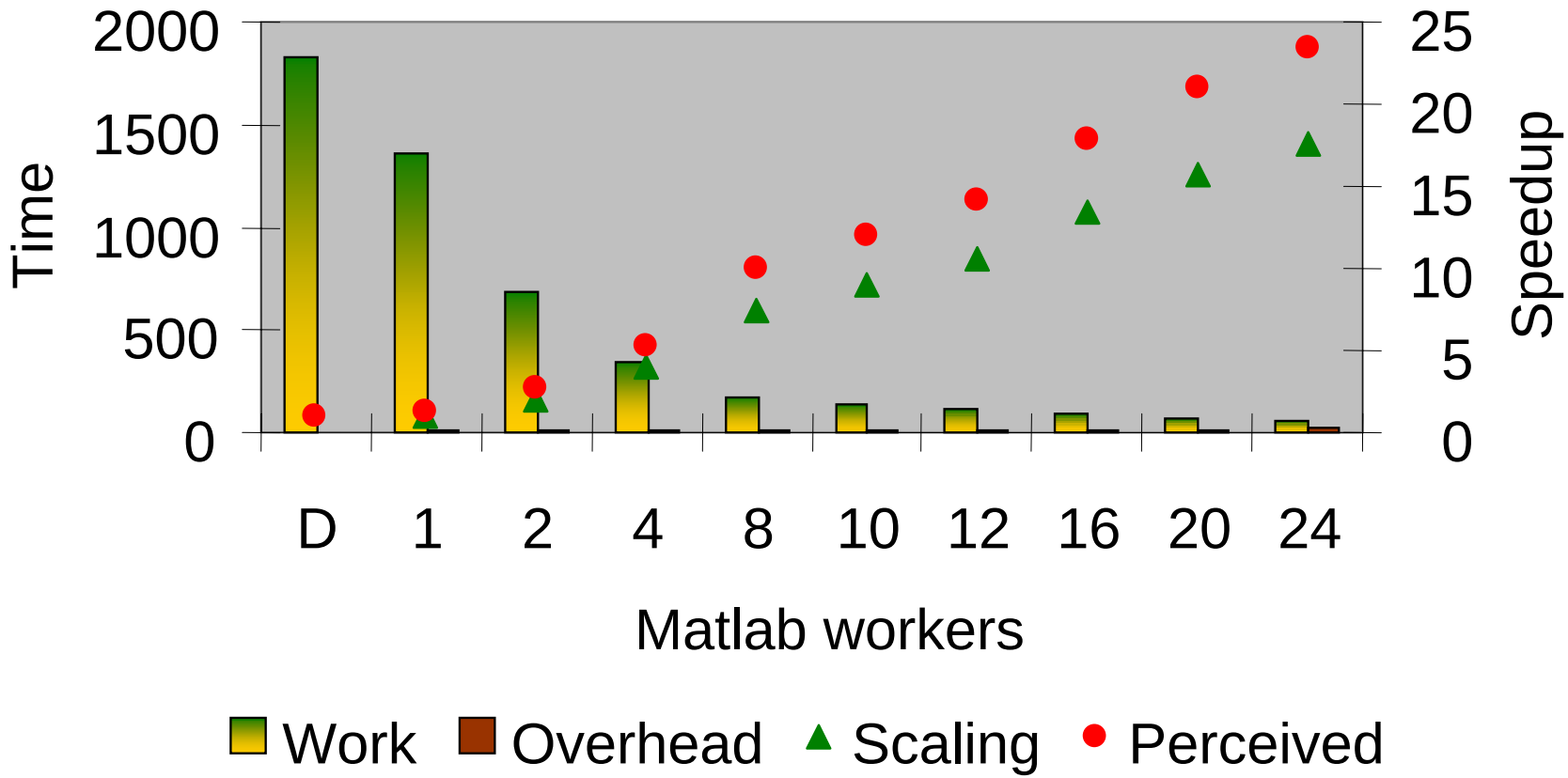
- Receptor is a moving target
- Fit a flexible ligand in the receptor by:
  - Isolating important receptor motion
  - Modeling both molecules as robots
  - Using methods planning from robotics
  - Scoring according to energy

For more information: http://www.cs.rice.edu/~kavraki

Neural Information Coding in the Primary Visual Cortex

- Christopher Rozell & Don Johnson
  - Electrical & Computer Engineering
- Hypothesis:
  - these neurons are forming what is known as a 'sparse code' for the visual scene
- “Very hard” signal processing problem
- Require simulation of a very large system of ODEs
- Embarrassingly parallel problem
  - but, interesting problems will require >25GB memory (multiply two large matrices)
- Solution implemented in matlab
- Parallelized using Matlab’s distributed compute engine

Matlab MPI & Scaling

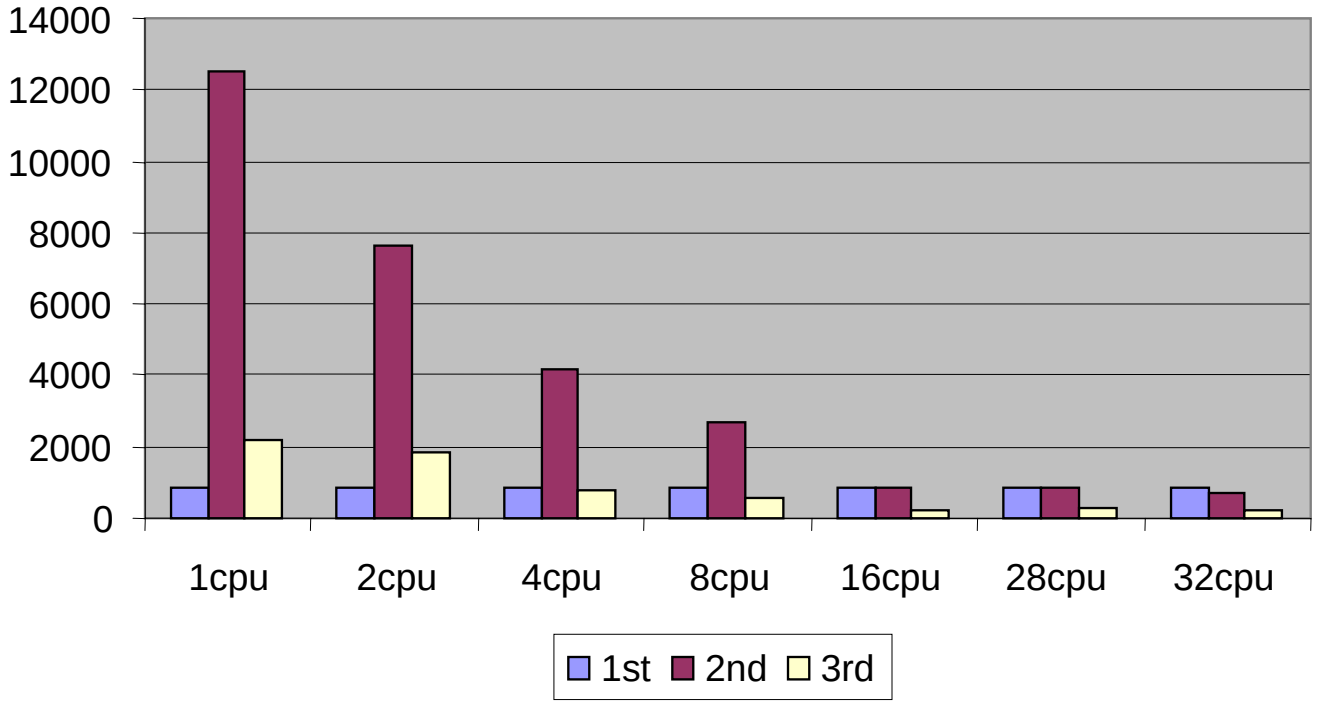


Polynomial Factoring

- C. Sidney Burrus, James Fox, Gary Sitton, and Sven Treitel
- **Lindsey-Fox algorithm**
  - uses the FFT (fast Fourier transform) to very efficiently conduct a grid search in the complex plane to find accurate approximations to the N roots (zeros) of an Nth degree polynomial
- Polynomials with real, random coefficients
- Algorithm for factoring high order polynomials (successfully factored >1M degree polynomial)
- Algorithm developed in matlab with C subroutines
- Factoring a 1M order polynomial takes literally days on a powerful desktop

Factoring order 400,000 polynomial

- Parallelized using Matlab's distributed compute engine



Open source: http://www.dsp.rice.edu/software/fvhdp.shtml

Parallelized version not posted yet

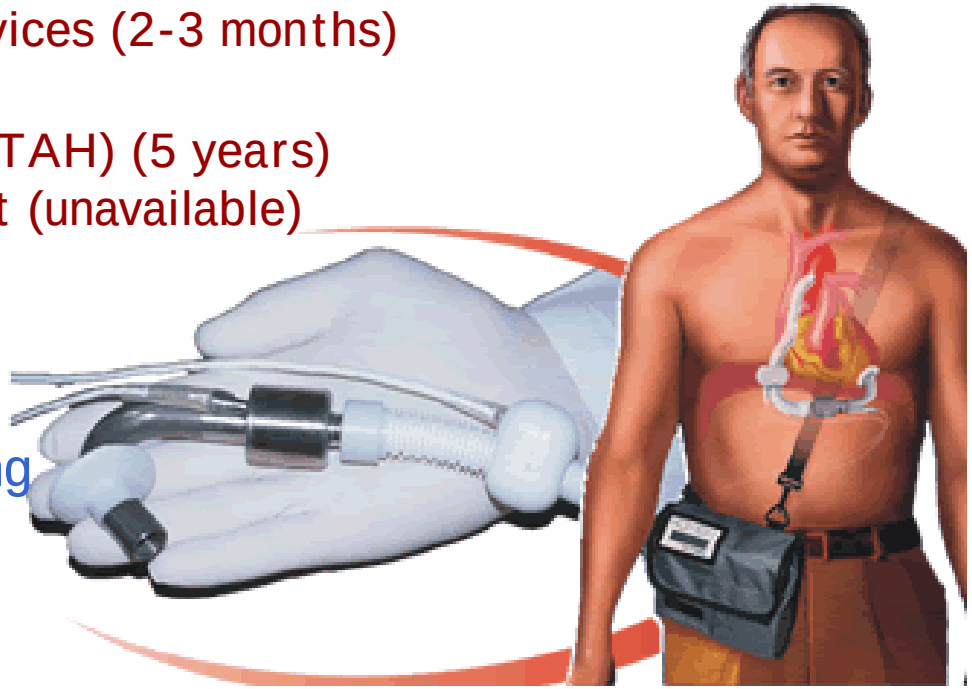
Flow in Micromed DeBakey Pump

Dhruv, Pasquali & Behr

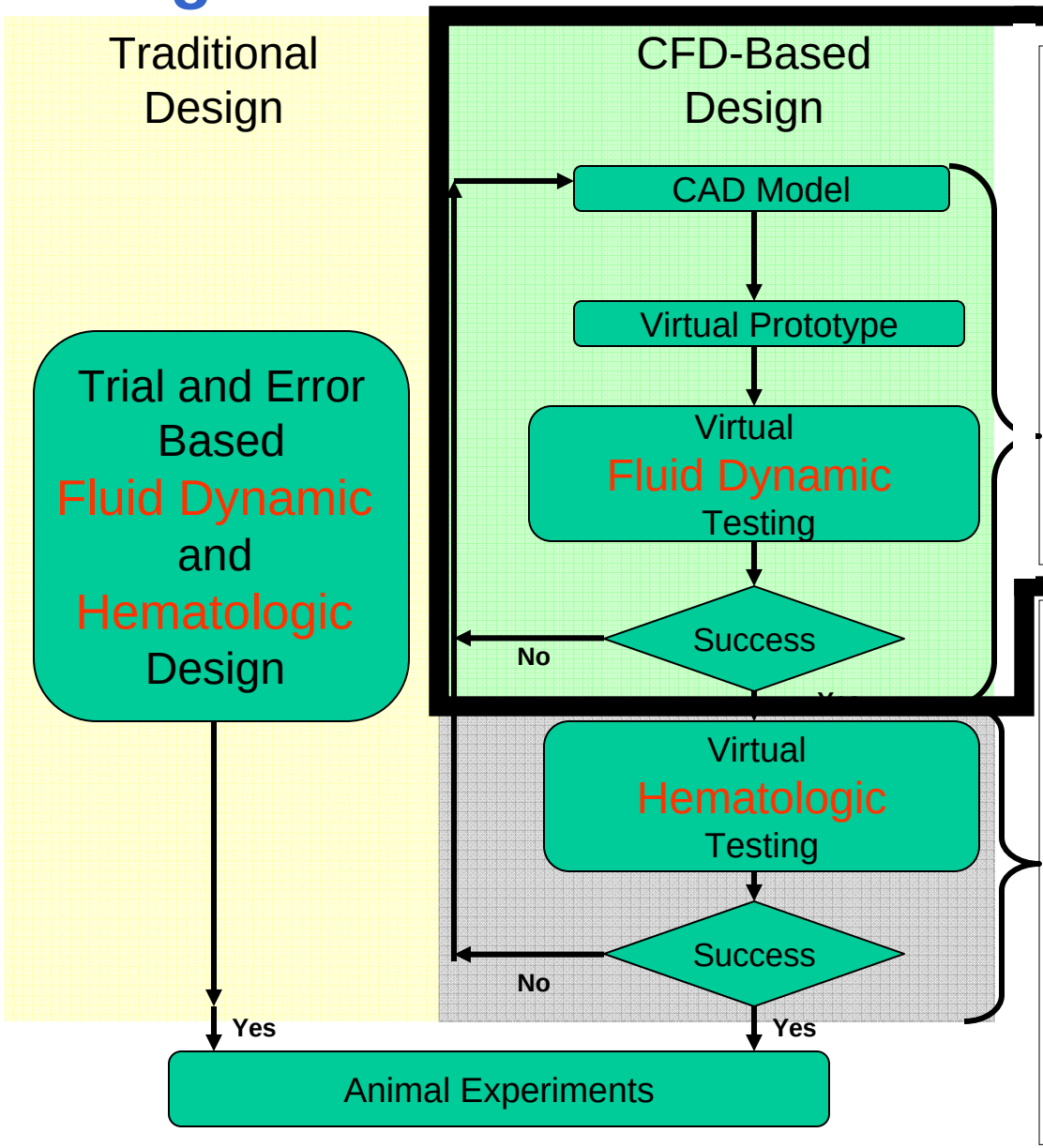
- Every year:
  - > 800,000 cases of heart disease are reported
  - > 50,000 patients need new hearts
  - > 2500 donor hearts available (5 %)

- Ventricular Assist Devices (VAD):
  - > Presently:
    - Bridge-to-transplant devices (2-3 months)
  - > In future:
    - Total Artificial Hearts (TAH) (5 years)
    - Alternative to transplant (unavailable)

- Design requirements
  - > Highly efficient
  - > Minimally blood-damaging



Design of VADs



Fluid Dynamic Design

Objective: To obtain appropriate flow rate and pressure head

- Governing equations
- Robust numerical methods
- Flow behavior known

Challenge: Simulate "blood flow"?

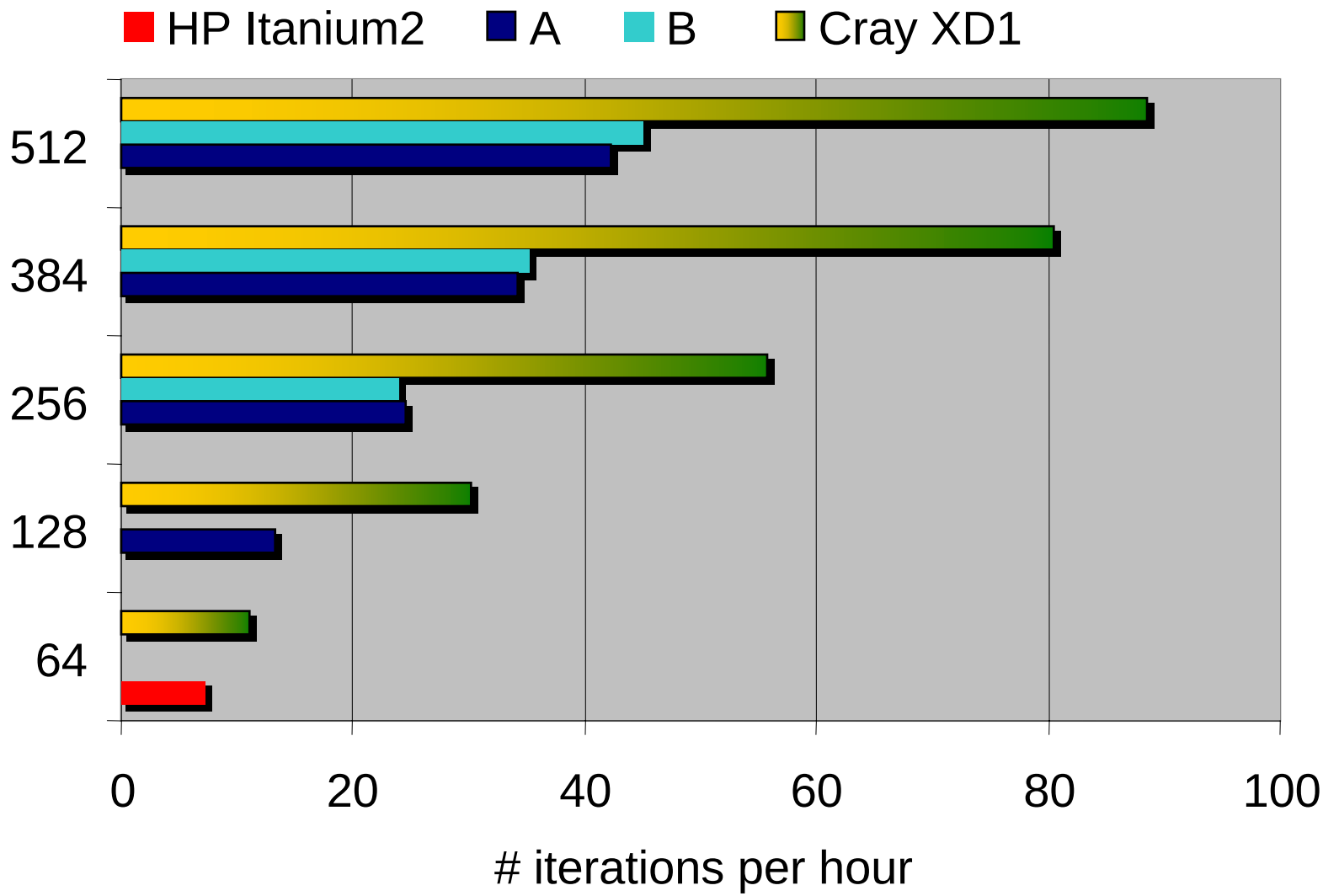
Hematologic Design

Objective: To compute blood damage

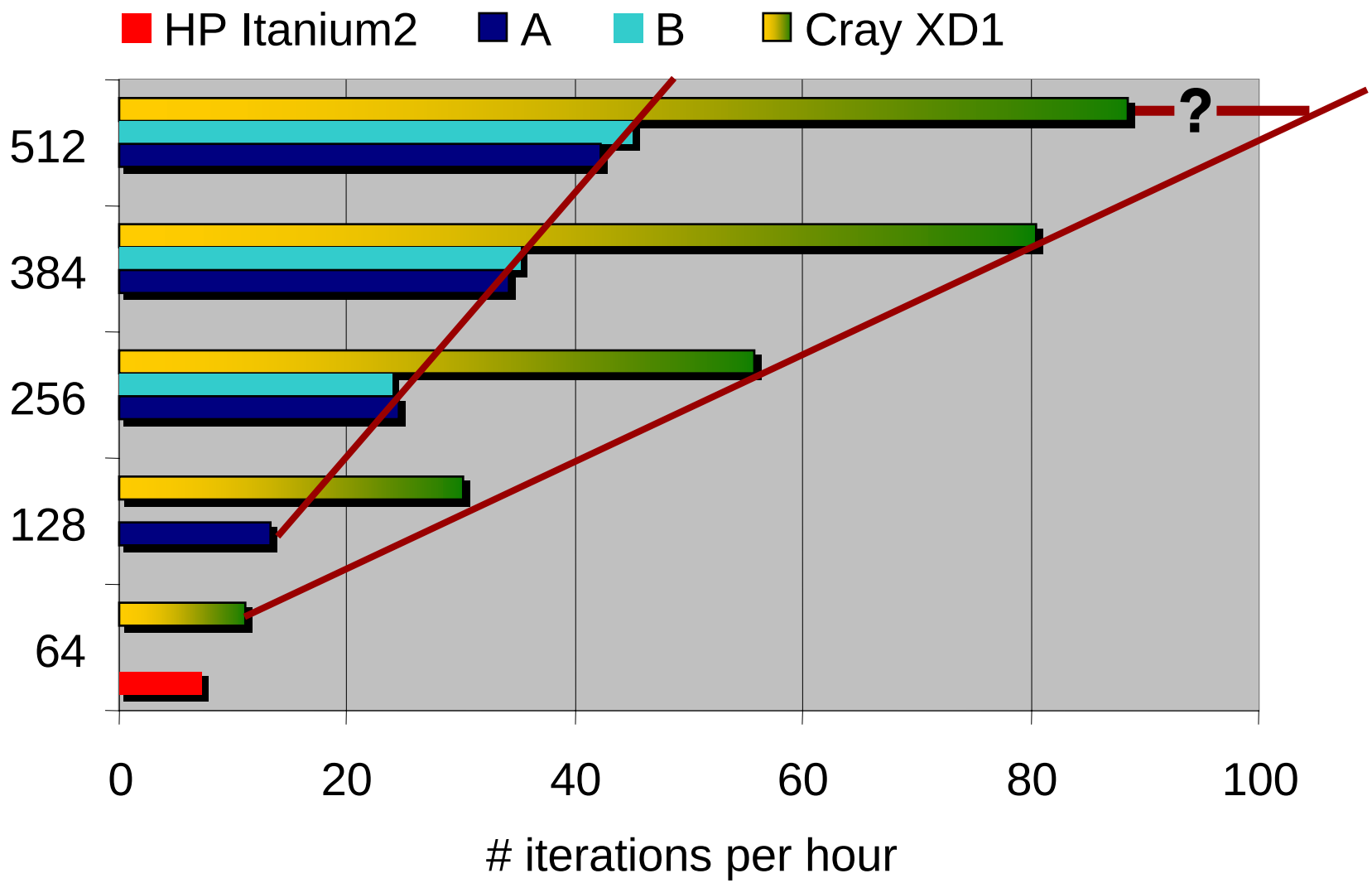
- Hemolysis
  - > Hemoglobin is toxic
- Thrombosis
- Not well understood

Challenge: Model blood-damage?

Performance & Scaling



Closing the Gap




Thank You