NSF Supercomputing: from the dark ages to Kraken

Phil Andrews,
Project Director, National Institute for Computational Sciences
University of Tennessee/Oak Ridge National Laboratory
andrewspl@utk.edu
Kraken:
1) legendary sea monster of gargantuan size
2) NSF Supercomputer at UT/ORNL
How did we get where we are now: National Science Foundation Supercomputing Timeline

• The Wilderness: pre 1984
• First Proposals: 1983-1984
• Centers Program: 1986-1996
• PACI Program: 1997-2004
• Office of Cyber Infrastructure: 2005
• TCS+DTF+ETF=TeraGrid: 2000-2010
• Track2a-d: 2007-2010
• TeraGrid XD 2010
• Track1: 2011
Lessons to be learned from the journey:

• Every decision has consequences
• Sometimes, the biggest effects come from unintended side-issues
• The capabilities available are truly massive
• Enormous compute facilities require enormous support: hardware and people
• You should always have Fun!
Various Influences:

• Funding Patterns
• Research-Production Tension (Computer Science-Applications Support)
• Technology Changes
• Center “Personalities”
NSF Funding Patterns

• NSF primarily funds smaller grants, normally “sunset” after a few years, not persistent infrastructure

• First approach: 10 years with 5 year “internal review”

• Longest grant is now 4 years with TeraGrid closest to persistent “glue”. New TG XD proposals submitted yesterday for 5 years funding
Computer Science Research V. Application Support

• Continual oscillation between “Computer Science” and “Production Services” dominance

• Period of ~5 years between opposite extremes

• Dampened by consistent personnel and funding (NSF and centers), amplified by personnel changes and new funding initiatives

• Last CS Max ~2001 → TeraGrid formation

• Last PS Max ~2005/6 → OCI, Track1 & Track2 RFPs
NSF Supercomputing Technology Changes over the Years:

- Initial systems were all Vector processors (Cray, ETA): favoured vectorizable code
- Current systems are all MPP or clusters (IBM, Dell, Sun, Cray): favour scalable code
- Seen SIMD (Intel, KSR), Multi-Threaded (MTA), etc.
- Future will probably bring accelerators, etc.
- Each technology decision tends to severely impact the application space
Early computers were by Artisans like Charles Babbage's Difference engine: 31 digits of accuracy, 3 tons, designed in 1812
Or Seymour Cray

Seymour Cray standing next to the core unit of the Cray 1 computer, circa 1974

Photograph courtesy of the Charles Babbage Institute, University of Minnesota, Minneapolis
NSF HPC: the Early Years

• Little “open” capability, NMF ECC for Fusion

• Early 1980's: Various studies of academic need for and access to supercomputers.

• (Lax Report, Bardon-Curtis Report, Press Report, FCCSET study) Program Advisory Committee at NSF, led by Neal Lane incorporating about 25-30 distinguished computational scientists meets frequently.
Early Years (2)

- October, 1983 Princeton unsolicited proposal for Supercomputer Center (Steve Orszag)
- November 1983: Unsolicited Proposal from U. Illinois (Larry Smarr)
- February 1984: Unsolicited Proposal from GA/UCSD (Sid Karin)
Early Years (3)

- Early 1985: Announcement of three center awards: SDSC, JvNC, NCSA, also announcement of an award in the new category of "Advanced Prototype" to Cornell.
- Cornell later became a center
- Late 1985: Friendly users at the centers.
- Jan 1, 1986: Official opening of SDSC, NCSA.
- 1986: Award to PSC
Initial Technologies (1986)

- NCSA (UI) and SDSC (GA): Cray XMP48
- JVNC (Princeton): award for ETA10, interim CDC 205 systems
- PSC (Pittsburgh): Cray 1S → Cray XMP48
- CTC (Cornell): Experimental hardware → IBM SP (much later)
- Staffing: centers at about 60-100 people
- Vector processing dominant
Initial Hardware (2)
Fig. 1-2. SDSC hardware configuration
NSF Centers
Organizational Changes

• Approximately 1986: CISE formed, OASC becomes DASC within CISE, Networking function separated out as DNCRI. Division level advisory committees are abolished.

• Some representation of computational science is maintained with one or two slots on CISE Advisory Committee.

• 1987: National Allocation committee formed for PSC and NCSA
Ancillary Activities: Networking

- NSFnet formed in 1986-7 (Dennis Jennings) to connect centers. NSF enforced adoption of TCP/IP standards for networking
- NCSA Telnet provided full TCP/IP stack for PC, Macintosh and Sun
- NSFnet would eventually transition to commodity internet (1995)
NSF Networking
Expanding NSFNet, late 80’s
How large-scale things evolve: From Arpanet to Internet
The Second Five Years

• Review in 1990 eliminated JVNC, renewed SDSC, NCSA, PSC, and CTC until 1995

• Confidence in funding lead to 3 major areas of expansion for centers:

  1) NSF MetaCenter
  
  2) Open Software
  
  3) Parallel Systems Technology
Ancillary Activities: Open Software Development

- NCSA Telnet: became model for “open source” tools.
- Gplot, Gdoc, & P3D from PSC, 1,100 sites
- SDSC Image Tools
- NCSA Mosaic: ultimate NSF Center software tool
- The Centers had developed development and support facilities for extensively used tools
GDOC


• Full SGML parser, used DTD for HTML, 90,000 lines of C++

• First browser with integrated graphics

• Talk and Installation at NCSA (’91)

• Funding for browsers hard to acquire (even with videotapes!)
NCSA Mosaic

- Mosaic: Mapping Cyberspace, Hardin, Grossman, Michaelson, Andreessen, and Andrews, submission to NSF, 1993
- Jim Clark (ex of SGI) hired Marc Andreessen and the coding team: rewrote Mosaic in a “clean room”. Great investment-> NetScape
- Started both the Internet IPO boom and the browser wars
Historical Plaque at U. Illinois!

WEB BROWSER

MOSAIC, THE FIRST POPULAR GRAPHICAL BROWSER FOR THE WORLD WIDE WEB, WAS CREATED BY MARC L. ANDREESEN AND ERIC J. BINA AT THE NATIONAL CENTER FOR SUPERCOMPUTING APPLICATIONS (NCSA). UPON ITS 1993 RELEASE TO THE PUBLIC, MOSAIC GAVE INTERNET USERS EASY ACCESS TO MULTIMEDIA SOURCES OF INFORMATION. WEB BROWSERS HAVE TRANSFORMED THE EXCHANGE OF INFORMATION.

UNIVERSITY OF ILLINOIS
Massively Parallel Processing (MPP)

- In 1995, 15 parallel processing systems within the centers
- Diverse architectures: Intel Paragon, nCUBE, TMC CM2, TMC CM5, Kendall Square, Workstation clusters from DEC, HP, IBM, IBM SP, Cray T3D, Convex, SGI Challengers
- Vector systems were still the “workhorses”, e.g., Cray C90 at PSC
The PACI Years


- In the winter of 1995, NSF issued a program solicitation, and called for a competition for "Partnerships for Advanced Computational Infrastructure" (PACI)

- Centers funding extended for 2 years, then 5+5 for PACI program
The Partnership for Advanced Computational Infrastructure

- Competition for partners was intense, with NCSA and SDSC incorporating over 75 sites
- “leading edge” sites were reduced to 2 with PSC and CTC not renewed
- Immediate shortfall of cycles without PSC
- SDSC acquired Cray T90 to compensate
NPACI Partners

- 46 institutions
- 20 states
- 4 countries
- 5 national labs
- Many projects
- Vendors and industry
- Government agencies
Changes in Supercomputer Center Design: We used to build world’s tallest buildings...
Now we build world’s longest roads...
Chip development cost by Generation

<table>
<thead>
<tr>
<th>Generation</th>
<th>Relative cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>8080</td>
<td>1</td>
</tr>
<tr>
<td>8086</td>
<td></td>
</tr>
<tr>
<td>80286</td>
<td></td>
</tr>
<tr>
<td>80386</td>
<td></td>
</tr>
<tr>
<td>80486</td>
<td></td>
</tr>
<tr>
<td>Pentium</td>
<td></td>
</tr>
<tr>
<td>P6</td>
<td></td>
</tr>
</tbody>
</table>

- Relative cost
  - 8080 = 1
The difference between then and now…

• Earlier Supercomputers were like tall buildings: monolithic and designed as a unit, rarely identical and difficult to change

• Now, Supercomputers are generally made up of smaller components: pieces can be added or taken away at any time

• The problems of modern HEC centers are those of civil engineering: weight, power, heat and cooling
Pontifications:

- The main reason behind the exponential growth in Moore’s law is the exponential growth in the money spent on it (corollary: will slow soon)

- In a free market (not Formula 1) HPC will only have mass market processors and memory available to it

- The single specificity to HPC will be the node interconnect because that is the only piece not yet needed by the mass market
New systems accumulate (like roads)
Distributed TeraScale Facility (DTF)

- 2001: NSF sent “Dear Colleagues” letter to SDSC, NCSA to provide coordinated systems
- For complex reasons, CACR (CalTech) and Argonne National Laboratory (ANL) added
- Significant funding earmarked for connecting network: initially intended as a mesh, then a more flexible “Grid”
TeraGrid WAN and Regional Networks

One Wilshire (Carrier Fiber Collocation Facility)

455 N. Cityfront Plaza (Qwest-operated Fiber Collocation Facility)

Los Angeles

Chicago

DTF Backbone Core Switch

Qwest POP at JPL

Abilene and other External Networks

Qwest San Diego POP

215mi

2200mi

115mi

2mi

20mi

15mi

2mi

20mi

2mi

NCSA

ANL

Caltech Cluster

SDSC Cluster

NCSA Cluster

ANL Cluster

Site Border Switch

Cluster Aggregation Switch

Caltech

SDSC

NCSA

ANL

Ciena Corestream

Long-Haul DWDM (Operated by site)

Metro DWDM (Operated by Qwest)

Vendor TBD

Metro DWDM (Operated by Qwest)

Ciena Corestream

Long-Haul DWDM (Operated by Qwest)

Cisco 6500 Switch/Router (256 Gb/s crossbar)

DTF Local Site Resources and External Network Connections

Wavelengths (lambdas)
DTF Technologies

- Homogenous IA64 clusters at all 4 sites
- GbE connectivity to every node
- 30 Gb/s to each of the 4 sites
- 40 Gb/s backbone Chicago-LA
- Intended to provide distributed MPI capability
- 4 orders of magnitude latency difference between machine room and WAN connectivity!
TeraGrid Network
NSF HPC Reorganization

• 2005: Supercomputing was removed from CISE and became the Office of CyberInfrastructure reporting to NSF Director

• Separate Research awards, SDCI and STCI, established

• Other directorates have strong input into directions of OCI: more service oriented, shorter, more prescriptive awards
Tracks 1 and 2, 2005: A New Beginning

• 4 yearly Track 2 RFPs announced: $30M for capital, operations funded separately, production in 2007, 8, 9, 10

• Single Track 1 RFP, $200M for “sustained PetaScale system”, production in 2011

• Very prescriptive on allowed work: very oriented to production work, very low operations money

• TeraGrid connectivity important
Track2A award

• Recommended to TACC (U.T. Austin) in 2006, awarded in 2007

• Sun System, AMD processors, CBBW network interconnect, custom (Magnum) switch (2)

• Over 60,000 cores, 500+ TF, ~125TB, ~3.4MW

• Began production Q1, 2008
Track2A, Texas Advanced Computing Center, Austin, Texas
<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track-2b Award Announced</td>
<td>Sep 2007</td>
</tr>
<tr>
<td>First Resources Available for Friendly Users</td>
<td>May 2008</td>
</tr>
<tr>
<td>Cray XT3 Production – 40 TF</td>
<td>Jun 2008</td>
</tr>
<tr>
<td>Cray XT4 Acceptance</td>
<td>Jul 2008</td>
</tr>
<tr>
<td>Cray XT4 Production 18K cores - 166 TF</td>
<td>18 Aug 2008</td>
</tr>
<tr>
<td>Cray XT5 Delivery</td>
<td>Oct-Dec 2008</td>
</tr>
<tr>
<td>Cray XT5 Production 64K cores – 615 TF</td>
<td>Feb 2009</td>
</tr>
<tr>
<td>Cray XT5 Upgrade ~1 PF</td>
<td>Dec 2009</td>
</tr>
</tbody>
</table>
Track2B, NICS, Oak Ridge, Tennessee
Challenges and Opportunities at NICS

- Staffing is the largest short-term challenge
- Helping researchers take advantage of multicore architectures
- Co-location with ORNL is a big advantage for NICS
  - Leverage: physical environment, cheap power, vendor presence, archival storage, experienced staff, joint training and outreach, existing documentation, visualization resources, …
- U of Tennessee is fertile ground for petascale projects, partnerships, and potential staff
- 20% of NICS resources targeted to outreach – nationally and in the State of Tennessee – academic and industry
  - Outreach goal: create new computational scientists
Track2C award to PSC/SGI

- Large number (>40 ?) Single System Images
- 2048 cores per SSI ?
- Low memory per core?
- ~1PF?
- Mixed mode programming necessary for out-of-box programming?
- Significant NUMA?
Track1 award

- Awarded to National Center for Supercomputing Applications (U.Illinois, Urbana-Champaign) in 2007
- IBM “PERCS” system (Power Series)
- Sustained PetaFlop
- >10 PF, 1 PB memory, 10PB disk, 1EB storage
- 100 Gb/s WAN connectivity
Track1, NCSA, U. Illinois
NICS Current Technology Status

• Kraken has been successfully upgraded from XT3 (39TF) to XT4 (167TF) and is in production

• Larger file system entering production this week, increasing gross disk from 100TB to 460TB

• Second round of allocations began Oct 1; new accounts have been created

• Many 16-18K core jobs already run (23% of total)
Cray XT4 Compute Module

- 4 AMD Opteron sockets
- 4 DDR2 DIMMs per socket. Support for 2, 4, 6 and 8 GB per socket
- >100 GB/sec Blade connector
- Redundant VRM
- L0 Blade Control Computer
- SeaStar2 Mezzanine Card
Facility Modifications

- Two 13.8kV/600A feeders increase NCCS electrical distribution capacity to more than 16MVA

- Two 1500-ton chillers added to the Central Energy Plant provide enough chilled water to manage another 10MW of heat load

- Two 480V switchboards provide up to 128 individual 100A breakers for Cray XT5 systems.

- PDUs provide redundant power (one side on rotary UPS) to file systems, HPSS, network components, and critical infrastructure servers.

Electricians work to position a 2500/3300kVA transformer, one of two that will provide 480V service to the NICS Cray XT5.
NCCS XT5 cabinets are here
3.3 PB of Disks for NICS XT5
Space is cleared for Kraken XT5
Kraken: Number of NSF Allocation Awards by Field of Science

- Molecular Biosciences; 20
- Astronomical Sciences; 7
- Advanced Scientific Computing; 6
- Atmospheric Sciences; 6
- Physics; 5
- Earth Sciences; 4
- Chemical, Thermal Systems; 4
- Materials Research; 2
- Software Systems; 1
- Elementary Particle Physics; 2
- Cross-Disciplinary Activities; 1
- Integrative Biology and Neuroscience; 1
- Mechanical and Structural Systems; 1
- Design and Manufacturing Systems; 1
- Environmental Biology; 1

Total Awards: 47
Kraken: Hours Awarded by Field of Science

- Atmospheric Sciences: 6,502,000
- Advanced Scientific Computing: 870,940
- Physics: 3,142,000
- Cross-Disciplinary Activities: 200,000
- Materials Research: 400,000
- Software Systems: 200,000
- Mechanical and Structural Systems: 75,000
- Environmental Biology: 200,000
- Integrative Biology and Neuroscience: 50,000
- Design and Manufacturing Systems: 30,000
- Elementary Particle Physics: 21,102,800
- Astronomical Sciences: 17,934,360
- Chemical, Thermal Systems: 12,930,000
- Molecular Biosciences: 11,893,680
- Earth Sciences: 11,130,000
- Chemistry: 10,416,000

Total Hours Awarded: 87,711,500
Enzo, Astrophysics: Galaxy Formation (Mike Norman/Robert Harkness)

- Developed hybrid approach for Enzo cosmology code using OpenMP/MPI, appears to work well on Kraken and DataStar, limited by memory bw/socket
- Begun a UPC implementation
- Running at 2048^3 model size
P.K Yeung: Mixing on a supercomputer

- **Objective:** To simulate finest details of mixing in high Reynolds number turbulence

- **Description:** Velocity fields at $4096^3$ resolution are prepared for simulations of mixing and dispersion at level of detail and precision never attained before. Code exhibits 99% weak scaling from $2048^3$ on 2K cores to $4098^3$ on 16K cores.

- **Graphics:** 3D color rendering of concentration field, combined with line trace showing a ramp-and-cliff structure.
Strong Scaling by P.K. Yeung (turbulence), Friday, Oct 17

$N^3 = 2048^3, 4096^3, 8192^3$

- Squares: Ranger (TACC)
- Stars: BlueGene (SDSC/IBM)
- Triangles: Cray XT4’s

Very good scaling on several architectures, e.g.: > 98%
strong scaling on BG and Ranger from $M = 16K$ to $32K$

Sensitive to proc. grid $M_1 \times M_2$
(depends on network topology)

- World record: $4096^3$ by Yokokawa et al. SC’02, Kaneda et al. 2003

- Current production runs at $4096^3$ at 16K (Ranger and Kraken)

- Excellent scalability and best performance on Kraken
P.K. Yeung’s mixing visualization
Core Scaling for Kraken, DataStar,

Cellulose NVE Benchmark AMBER v10 408K Atoms

Throughput (NS/day)

Processors (MPI Threads)

Kraken 1ppn
Kraken 2ppn
Kraken 4ppn
Ranger 4ppn
Ranger 8ppn
Ranger 16ppn
DataStar 4ppn
DataStar 8ppn
DataStar 10ppn
Kraken 25% / SU
Kraken 50% / SU
DataStar 50% / SU
DataStar 100% / SU
Kraken 100% / SU
Ranger 25% / SU
Ranger 50% / SU
Ranger 100% / SU
HMMER – Protein Domain Identification tool
• existing MPI-HMMER – limited performance, did not scale well (see Fig 3. below)
• new HSP-HMMER – excellent performance (~100x faster than MPI-HMMER for 4096 cores) and scales well beyond 8000 cores (Figs. 4 & 5)
• HSP-HMMER code brings down time to identify functional domains in millions of proteins from 2 months down to less than 20 hours.
• HSP-HMMER paper accepted for publication in: ACM SAC 2009 Bioinformatics Track

Figure 3. Performance of MPI-HMMER on XT4.

Figure 4. Performance of MPI-HMMER and HSP-HMMER on XT4.

Figure 5. Multi-Threaded performance comparison of HSP-HMMER on XT4.
Greg Voth (Utah) on Kraken. Membrane Remodeling: A critical life-sustaining function that is not well understood.

Statistical Mechanics
\[ \chi^2(\phi) = \frac{1}{3N} \left( \sum_{i=1}^{N} \left| F_i^{MS}(M_R(r_i^n); \phi) - f_i(r_i^n) \right|^2 \right) \]

Coarse-grain (CG) molecular dynamics
\[ F_{ij}^{MS}(R_i, R_j, \phi) = \sum_{d=1}^{N_d} \phi_d \delta_C(R_d - R_{ij}) \hat{R}_{ij} \]

Biological Relevance

A 200 nm diameter liposome simulation on Kraken over 512 processors

Cryo-EM image of membrane remodeling via BAR domains, Vinzenz Unger, Yale.

Membrane remodeling at the atomic level.

The aim is to employ rigorous statistical mechanics, massively parallel CG simulation, and multiscale simulation to examine membrane remodeling.
Massively Parallel Systematic Coarse-grain (CG) Molecular Dynamics on Kraken

CG BAR domains: atomistic elastic network

CG 200 nm diameter liposome atomistic systematic CG

Largest CG membrane simulation performed: half a million lipids, over 400 BAR domains. Equivalent to over $10^{10}$ atoms.
All atom molecular dynamics simulations of the Arp2/3 branch junction:

- Fundamental investigation of structure/dynamics of key component of the cytoskeleton.
- Simulation currently running on Kraken has over 2.7 million atoms, one of the largest all-atom simulations of biomolecular systems.
- Scales linearly up to 4000 cores on Kraken.
- NICS provided support for compiling special versions of required software. Simulation not possible without NICS support.

Function: cleaves polyproteins during viral assembly

AIDS drug target

Remote mutations interfere with inhibitor binding

How can we connect genomics with structure and overcome drug resistance?
Adrian Roitberg using Amber on Kraken:
Study of the Conformations of Exenatide (exendin-4)

- Exenatide is a peptide of 39 amino acids.
- Currently commercialized as a treatment for type-2 diabetes. BYETTA (Amylin Pharmaceuticals, Inc. and Eli Lilly and Company)
- The high flexibility of this peptide produces challenges in both experimental and theoretical studies.
- Kraken will allow University of Florida researchers to study this molecule's conformational space using advanced molecular dynamics methods.
- This research will provide valuable information for the development of more effective treatments for diabetes.
MD simulations of Nickel regulatory protein (NikR)

- NIH grant # GM 066689 to Kenneth M. Merz Jr. in collaboration with Adrian E. Roitberg.
- Study of the Nickel-bound form of the Nickel regulatory protein derived from the Pyrococcus Horikoshii species of NikR originally crystallized by Peter Chivers and Tahir Tahirov in 2005.
- Novel all-atom model of this 552 amino acid residue long homo-tetramer.
- The dynamics revealed by this study will give unique insight into the allostery of not only over 30 species of bacteria and archaea, but also many similar metalloproteins found in humans.
UT IGERT (Keffer) Applications running on Kraken in October, 2008

- Flow of complex fluids
- Interfacial transport in nanostructured systems
- Reactive molecular dynamics
- Metal organic frameworks as explosive sensors
- Precise physical property determination (wet chemistry lab replacement)
Center for Integrated Space Weather Modeling, John Lyon, Dartmouth

- NSF Science and Technology Center
  - Space Weather: The effects of the Solar Wind and radiation upon the Earth’s space environment
  - Headquartered at Boston University – 9 Member Institutions
  - Goal is to create a set of coupled, front line simulations that can model phenomena stating on the Sun all the way to the Earth’s upper atmosphere
  - Coupling strategy
    - Separate executables linked by InterComm communications library and coupling modules.
    - InterComm handles MxN communications, grid friendly (can communicate between hardware platforms as well as one machine)
    - Coupling modules use Overture framework to interpolate between grids and do variable translation
    - Next slide shows suite of CISM Geospace Models
Spatial scales of codes indicated alongside:

- LFM MHD code - Dartmouth
- Magnetosphere
- 0.05/6 $R_E$
- ITM
- TIEGCM - NCAR
- 0.1/300 $R_E$
- Magnetic Reconnection
- 10^{-5}/1 $R_E$
- Rice Convection Model – Rice U
- 0.01/30 $R_E$
- Radiation Belts & SEP Entry
- 0.001/12 $R_E$
- MI Coupling
- 0.01/30 $R_E$
- Colorado, Dartmouth, Rice
Reflections/Lessons learned

• “Change is the only constant”

• Technology decisions necessarily impact applications

• Trust between funding agencies and awardees is essential

• Personalities strongly influence directions
Prognostications(1):

- Technology choices $\rightarrow$ Application restrictions
- Facilities and power $\rightarrow$ Technology restrictions
- Need at least 3-5 MW for current HPC planning
- Cu is expensive, need to plan ahead in relation to power infrastructure

Alternatively, close ties to major HPC facilities, plus adept programming techniques and status as a “feeder” site for larger sites can be cost-effective. Everyone wants to look good!
Prognostications(2):

- Moore’s law has gone back to its roots: more transistors rather than clock speed
- Cores (and flops) are becoming cheaper faster than memory and interconnects
- Memory per core will start to drop (2GB/core peak already passed?)
- Constant Bisection Bandwidth interconnects may be too expensive to maintain
- Memory and Bandwidth may be better way to rank systems
Prognostications(3):

• Current programming models seem inadequate for next capability systems (~100K+ cores)
• Education, porting, takes a long time!
• New models need to be developed, taught, used, before rapidly depreciating HPC systems appear
• Resistance to new programming models/languages is intense: each researcher only codes for a few years (between UG and Tenure!) Needs very strong commitment from vendors and multiple HPC centers/countries
What does this mean for you?

• Enormous capabilities arising
• Few researchers ready to take advantage
• Discretionary time available for Outreach and/or Tennessee researchers
• Applications should be designed with idea they may run on > $10^5$ cores!
• User Support: Bruce Loftis [bloftis1@utk.edu]