High-Performance Computing in the Scientific Community: Partnering with Industry
Seymour Cray founded Cray Research in 1972
- SGI purchased Cray Research in 1996

Cray Inc. formed April 2000
- Tera purchased Cray Research assets from SGI
- Nasdaq: CRAY
- 850 employees across 20 countries
- Headquarters in Seattle, WA

Five Major Development Sites:
- Chippewa Falls, WI
- St. Paul, MN
- Seattle, WA
- Austin, TX
- Pleasanton, CA
Redefining Supercomputing

We build solutions to help solve “Grand Challenges” in science, engineering and knowledge discovery

HPC Systems
Breakthrough performance and scalability

Storage & Data Management
Scalable, manageable high performance storage systems

Big Data Solutions
Large scale relationship analytics
Cray Specializes in Large Systems

Over 35 PF’s & $600+M In XE6 and XK7 Systems
Cray Industry Solutions

Industry-specific solutions that leverage Cray expertise across all technologies and products
Concurrency is growing

- Not just from the number of cores
- Longer vector lengths
  - Probably fixed length
- Systems made from different types of “processors”
  - Hybrids of scalar and vectors
- Requires new methods of programming to fully exploit the hardware

Average Number of Processor Cores per Supercomputer (Top20 of Top500)

Source: [www.top500.org](http://www.top500.org)
Key Challenges to Get to Exascale

**Power**
- Traditional voltage scaling is over
- Power now a major design constraint
- Cost of ownership
- Driving significant changes in architecture

**Concurrency**
- A billion operations per clock
- Billions of refs in flight at all times
- Will require *huge* problems
- Need to exploit *all* available parallelism

**Programming Difficulty**
- Concurrency and new micro-architectures will significantly complicate software
- Need to hide this complexity from the users

**Resiliency**
- Many more components
- Components getting less reliable
- Checkpoint bandwidth not scaling
- Impacts both systems and storage
Which Takes More Energy?

Performing a 64-bit floating-point FMA:

\[
\begin{align*}
893,500.288914668 & \times 43.90230564772498 \\
= 39,226,722.78026233027699 \\
+ 2.02789331400154 \\
= 39,226,724.80815564
\end{align*}
\]

This one takes over 3x the energy!

And loading the data from off chip takes > 10x more yet

- Flops are cheap, communication is expensive
- Exploiting data locality is critical for energy efficiency
What Will an Exascale System Look Like?

...no one knows exactly.

There are some macro level trends that provide a good indication.

Best approach is to prepare for:

- Massive parallelism
- Locality
- Minimized resource usage (eg: memory and operating system)
- Greater degree of complexity exposed to programmer and user

With regard to requirements, seek open and common solutions:

- From M/I/S/C article on dealing with complexity systems: “If everything in your environment is special, you are doing it wrong.”

This has been true for many years, but tolerance threshold at exascale will be much harsher.
Programming Models

**Current**

**MPI:**
- standard approach for distributed memory systems;
- provides no support for other architectural characteristics

**OpenMP:**
- focus on shared memory and loop parallelism;

**Hybrid (MPI + OpenMP):**
- Common for extending the maximum number usable cores
- adds some complexity to the program

**Higher level approaches – Matlab:**
- applicable to codes if bulk of computation can utilize optimized third party code;
- can provide portability
- Performance ??

**Emerging Technologies**

**Partitioned Global Array Space (PGAS) -**

**UPC, CAF, Titanium:**
- for distributed memory system -> explicit data “movement” using a global index space;
- solves one issue, no support for other challenges
- Optimally requires hardware support

**Specialized – CUDA, … :**
- some are non-portable due to focus on specialized hardware,
- others are not proven for large applications

**High Productivity Computing Systems (HPCS) - X10, Chapel, Fortress:**
- focus on productivity
- relying on compiler technology to enhance performance
Today HPC relies on hybrid-computing to achieve scale and sustained performance per $/\text{watt}$, must adopt:

- **Hybrid node architecture**
  - Fast cores for serial code
  - Many power-efficient cores for parallel code

- **Deep, explicitly managed memory hierarchy**
  - Better exploit locality, improve predictability, and reduce overhead

- **Microarchitecture to exploit parallelism at all levels of a code**
  - Distributed memory, shared memory, vector/SIMD, multithreaded

Back to the Future: Impact on Applications Developers

Hardware

- For the next decade all HPC system will basically have the same architecture
  - Message passing between nodes
  - Multi-threading within the node
  - Vectorization at the lower level

Applications

- Current petascale applications are not structured to take advantage of these architectures
  - Current – 80-90% of application use a single level of parallelism, message passing between the cores of the MPP system
  - Looking forward, application developers are faced with a significant task in preparing their applications for the future

---

Larkin, J and Levesque, J, Application Development for the Cray XK6, CUG 2012

Why Accelerated Computing?

The great technology inflection point of mid last decade:
- **Power** has become a huge cost and the major design constraint
- Future **performance from parallelism**. Not clock rate. Not complexity
- Targeting a 100x improvement in sustained flops/W this decade!

**Multi-core x86 was a great first response…**
- Performance through parallel cores; exploit locality; modest clock rate
- …but it won’t get us where we need to go:
  - Standard cores are designed to run threads fast, not power-efficiently

**Conclusion:**
- Need lots of very simple cores optimized for power on parallel code
  - But also need some fast cores for serial code
- **Heterogeneous processing** (a.k.a. accelerators)

**GPUs are looking very promising for HPC**
- They now have good Double Precision Floating Point and Error Protection
- They have high volume!
  - ship O(100M) units a year
- It looks like they can credibly support both masters (graphics and compute)
- **Big issues** are **structural bottlenecks** and **programming difficulties**
Heterogeneous Multi Petaflop Systems

Blue Waters: Sustained Petascale Performance
Production Science at Full Scale
235 XE Cabinets + 30 XK Cabinets
  - > 25K compute nodes
11.5 Petaflops
1.5 Petabytes of total memory
25 Petabytes Storage
  - 1 TB/sec IO
Cray’s scalable Linux Environment
HPC-focused GPU/CPU Programming Environment

Titan: A “Jaguar-Size” System with GPUs
200 cabinets
18,688 compute nodes
25x32x24 3D torus (22.5 TB/s global BW)
128 I/O blades (512 PCIe-2 @ 16 GB/s bidir)
1,278 TB of memory
4,352 sq. ft.
10 MW
Current Hardware with and without Accelerators
Energy's Oak Ridge National Laboratory (ORNL). Titan is capable of more than 20 petaflops of high performance computing (HPC) power and is the world’s most powerful supercomputer for open science.

The Titan system is a 200-cabinet Cray XK7 supercomputer with 18,688 compute nodes each consisting of a 16-Core AMD Opteron™ 6200 Series processor and an NVIDIA® Tesla® K20 GPU Accelerator. Titan was upgraded from a Cray XT5 supercomputer nicknamed "Jaguar."

http://blogs.nvidia.com/2012/05/gtc-session-on-titan-supercomputer-showcases-science-on-gpus/
Cray XK6 -- Hybrid System

Components

- AMD Series 6200 16-core Interlagos processors
- NVIDIA® Tesla™ 20-series many-core processors
- Cray Gemini High-Performance Interconnect
- CLE and CPE, Cray’s scalable software environment

Hybrid Supercomputer

Combining CPU technology with GPU technology
Cray XE6/XK7 System Architecture

- GigE
- Login Server(s)
- Network(s)
- 10 GigE
- Infiniband
- Fibre Channel
- GigE
- SMW
- RAID Subsystem(s)
- Network Attached Storage

Service Nodes:
- Login/Network
  - Login
  - Network
- Lustre File System
  - Metadata Server
  - Object Storage Server
- Operating System
  - Boot
  - System Database
- Cray DVS
  - DVS Server
System balance at scale requires a much more tightly integrated architecture...especially as node/socket performance grows.
## XK7 Compute Node Characteristics

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host Processor</td>
<td>AMD Series 6200 (Interlagos)</td>
</tr>
<tr>
<td>Host Processor Perf.</td>
<td>90-147 Gflops</td>
</tr>
<tr>
<td>Tesla K20 Cores</td>
<td>2496</td>
</tr>
<tr>
<td>Tesla K20 Perf.</td>
<td>1.17 Tflop/s</td>
</tr>
<tr>
<td>Host Memory</td>
<td>16 or 32GB 1600 MHz DDR3</td>
</tr>
<tr>
<td>Tesla K20 Memory</td>
<td>5 GB GDDR5 capacity 200 GB/sec</td>
</tr>
<tr>
<td>Gemini Interconnect</td>
<td></td>
</tr>
<tr>
<td>Dynamic Parallelism</td>
<td></td>
</tr>
<tr>
<td>Hyper-Q</td>
<td></td>
</tr>
</tbody>
</table>
Cray XK7 Blade

- Cray Gemini
- AMD 64
- NVIDIA
- X2090 GPU per node
- Four Compute Nodes
Adaptive Software for the Cray XK6 and XE6

Cray Software Ecosystem for XE6 – now extended to XK6

Cray Programming Environment (CPE) Extensions

• Now an adaptive GPU/x86 Software Environment

- PGI for x86/GPUs
- CUDA NVIDIA SDK
- Enhanced Libraries
- Cray Compiler for x86/GPUs
- Cray Optimization Tools for x86/GPUs
- Debuggers 3rd Party Tools
Programming Models

OpenACC®
DIRECTIVES FOR ACCELERATORS

http://www.openacc-standard.org/

NVIDIA


Khronos Group
CONNECTING SOFTWARE TO SILICON

http://www.khronos.org/opencl/
Fortran, C, and C++ compilers

- **Directives to drive compiler optimization**
  - Compiler does the “heavy lifting” to split off the work destined for the accelerator and perform the necessary data transfers
  - Compiler optimizations to take advantage of accelerator and multi-core X86 hardware appropriately
- Advanced users can mix CUDA functions with compiler-generated accelerator code
- Debugger support with DDT

Cray Reveal, built upon an internal compiler database containing a representation of the application (the CCE Program Library)

- Source code browsing tool that provides interface between the user, the compiler, and the performance analysis tool
  - **Scoping tool** to help users port and optimize applications
  - **Performance measurement and analysis** information for porting and optimization

Scientific Libraries support

- Auto-tuned libraries (using Cray Auto-Tuning Framework)
Scaling (running big jobs with a large number of GPUs)

- **Results** summarized and **consolidated** in one place

**Statistics for the whole application**

- Performance **statistics mapped back** the user **source** by line number
- Performance statistics grouped by accelerator directive
- Single report can include **statistics for both the host and the accelerator**

**Single tool for GPU and CPU performance analysis**

- Performance statistics
  - Includes accelerator time, host time, and amount of data copied to/from the accelerator
- Kernel level statistics
- Accelerator hardware counters
New code restructuring and analysis assistant...

- Uses both the performance toolset and CCE’s program library functionality to provide static and runtime analysis information and hints.
- Assists user with the code optimization phase by correlating source code with analysis to help identify which areas are key candidates for optimization.
New code restructuring and analysis assistant...

- Uses both the performance toolset and CCE’s program library functionality to provide static and runtime analysis information

- Assists user with the code optimization phase by correlating source code with analysis to help identify which areas are key candidates for optimization

Key Features

- **Annotated source code with compiler optimization information**
  - Provides feedback on critical dependencies that prevent optimizations

- **Scoping analysis**
  - Identifies shared, private and ambiguous arrays
  - Allows user to privatize ambiguous arrays
  - Allows user to override dependency analysis

- **Source code navigation**
  - Uses performance data collected through CrayPat
Visualize CCE’s Loopmark with Performance Profile
Industry-Academic Partnership

“If you ask me, between your stuff and my stuff, there's just nothing left to invent.”

Source: Nature Chemical Biology Volume 2 Number 12  December 2006 pp.649
Blue Waters Project

Supercomputing at Sustained Petascale Performance

- $188M Contract
- Sustained Petascale Performance
- Cray Sonexion Storage
- 11.5PF, 25PB, 1 TB/sec IO
- Production Science at Full Scale
- Cray Reliability and Service
Selected Petascale Computing Resource Allocations winners collaborating with NCSA and the Blue Waters team include:
[http://www.ncsa.illinois.edu/BlueWaters/prac-teams.html]

- Simulation of contagion on very large social networks with Blue Waters
  Principal Investigators: Keith Bisset, Virginia Tech; Shawn Brown, Carnegie-Mellon University; Douglas Roberts, Research Triangle Institute
- The computational microscope
  Principal Investigator: Klaus Schulten, University of Illinois at Urbana-Champaign
- Computational chemistry at the petascale
  Principal Investigator: Monica Lamm, Iowa State University
- Simulating vesicle fusion on Blue Waters
  Principal Investigators: Vijay Pande, Stanford
- Hierarchical molecular dynamics sampling for assessing pathways and free energies of RNA catalysis, ligand binding, and conformational change
  Principal Investigators: Thomas Cheatham, University of Utah; Darrin York, Rutgers; Carlos Simmerling, State University of New York at Stony Brook; Adrian Roitberg, University of Florida; Ross Walker, San Diego Supercomputer Center
- Petascale multiscale simulations of biomolecular systems
  Principal Investigators: Gregory Voth and Benoit Roux, University of Chicago
- Petascale simulations of complex biological behavior in fluctuating environments
  Principal Investigator: Ilias Tagkopoulos, University of California-Davis
- Breakthrough petascale quantum Monte Carlo calculations
  Principal Investigator: Shiwei Zhang, College of William and Mary
Partnering with Developers: NAMD – Scalable Molecular Dynamics on Cray Supercomputers

http://www.ks.uiuc.edu/Research/namd/
Enabling and Scaling Biomolecular Simulations with 100 Million Atoms

NAMD
Scalable Molecular Dynamics

Atom Count

10^9
10^8
10^7
10^6
10^5
10^4


Lysozyme Apoa1 ATP Synthase STMV Ribosome Chromatophore

11/9/2012
"Biomolecular simulations are highly challenging to efficiently scale to a large number of processors. Blue Waters - a Cray supercomputer - will provide the resources to routinely run 100-million-atom biomolecular simulations with NAMD. This type of calculation was unthinkable just a few years ago."

Prof. Klaus Schulten, University of Illinois
Cray Working with NAMD Developers

*Leveraging Cray Gemini custom interconnect to scale to thousands of processors*

NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems.

Based on Charm++ parallel objects

NAMD scales to hundreds of processors on high-end parallel platforms and tens of processors on commodity clusters using gigabit ethernet

NAMD uses the popular molecular graphics program VMD for simulation setup and trajectory analysis, but is also file-compatible with AMBER, CHARMM, and X-PLOR

NAMD is distributed free of charge with source code
uGNI-based NAMD Version for Cray Gemini

Charmm++ asynchronous message-driven programming model leverages Cray Gemini interconnect

Optimization techniques

- Persistent messages
- Memory pool
- Intra-node communication

L. V. Kale, et al., A uGNI-Based Asynchronous Message-driven Runtime System for Cray Supercomputers with Gemini Interconnect [IPDPS 2012]

L. V. Kale, et al. SC’12 November 10-16, 2012, Salt Lake City, Utah
How applications are going to look like when exascale becomes a reality?

What do we need to do to leverage exascale architectures?
100-Million Atom Simulation

Source: Jim Phillips, SC'12
● NAMD is routinely being optimized as part of the NCSA Blue Waters project
● Charm++ leverages MPI-uGNI, NAMD 2.9 (new version) uses uGNI native
  ● uGNI-based NAMD version outperforms MPI-based version by 18% (Kale, et al.)
  ● SMP version leverages Cray multicore-nodes
● NAMD 2.9 tested on 100 Million atoms system on Cray (Kale, et al.)

L. V. Kale, et al. SC’12 November 10-16, 2012, Salt Lake City, Utah
NAMD 2.9, apoa1

- **Interlagos (IL) 2.1 GHz**: 8% Speedup
- **Interlagos (IL) 2.3 GHz**: 8% Speedup
- **Interlagos (IL) 2.5 GHz**: 16% Speedup
- **AbuDhabi (AD) 2.5 GHz**: 20% Speedup

The graph shows the relationship between the number of cores and the time per step for different frequencies and locations.
“The new system will be unique among supercomputers owned by U.S. universities. The new Cray will have more than 21,000 processor cores, high-performance GPUs, and an extremely fast Cray Gemini Interconnect. It will include the newest NVIDIA GPU accelerators and the latest AMD processors with 16 cores each. Big Red II represents a substantial capability leap forward for IU over Big Red's 4,100 cores. ”
The Era of Big Data

Source: Eric Green, Director, National Institute of Health: NextGen 101 Workshop
SAL: Sequencing Analysis Library

Next Generation Sequence Analysis

ICiS Summer Session 3A:
July 28 to August 4, 2012
Park City, Utah

http://www.icis.anl.gov/programs/summer2012-3a
Dr. Barry C. Bolding

Barry Bolding serves as Vice President, Storage & Data Management and Corporate Marketing, responsible for Cray’s storage solutions business and corporate marketing. Dr. Bolding was appointed Vice President in 2009, overseeing product management, as well as corporate and product marketing for Cray’s high performance computing solutions. Prior to 2009, he served as Cray’s Director of Product Marketing where he analyzed future products and developed long-term strategies. Over the course of his career, Dr. Bolding has worked with key customers in government, academia and commercial markets and held positions as a scientist, applications specialist, systems architect and presales product and marketing manager. He first joined Cray Research, Inc. in 1992 and has held subsequent positions with Network Computing Services and IBM. Dr. Bolding holds a B.S. in chemistry from the University of California at Davis and a Ph.D. in chemical physics from Stanford University.

http://genomedata-analysis.com/speakers
Life Sciences Analytics

The YarcData uRiKA graph appliance is a purpose built solution for Big Data relationship analytics.

uRiKA enables Life Scientists to:

- Discover unknown and hidden relationships in Big Data: uRiKA enables the building of a relationship warehouse, supporting inferencing/deduction, pattern-based queries and intuitive visualization
- Perform real-time analytics on Big Data graph problems: uRiKA is a high performance graph appliance with a large shared memory and massively multithreaded custom processor designed for graph processing and scalable I/O
- Realize rapid time to value on Big Data solutions: uRiKA is easy to adopt by Life Scientists with an industry standard, open source software stack enabling reuse of existing skill sets and no lock in
uRiKA Customer Use Case: Cancer Research Institute

The Challenge
- Multiple massive datasets describing biological network graphs in cancer cells from published literature and experimental data, constantly updated
- Un-partitionable, densely and irregularly connected graphs
- Multiple researchers concurrently searching for relationships not found in published literature

uRiKA Solution
- uRiKA holds un-partitioned genome and protein family network graph in memory
- Contrast experimental models and theories with published results to discover previously unknown relationships
- Interactive, real time access by multiple researchers

Business Value
- Identify new pathways in cell models to refine cancer treatments

Confirmation of elevated VEGF levels by tissue microarray:
Summary

Cray works with scientists and engineers to leverage Cray systems to develop solutions to solve their critical problems on:

- Large simulations on the world’s fastest supercomputers
- Mid-range production systems as a dedicated or shared-resources for groups working in Engineering and Scientific Computing
- Genomic large data challenges
- YarcData solution enables the building of a relationship warehouse, supporting inferencing/deduction, pattern-based queries and intuitive visualization
"The future is seldom the same as the past"

Seymour Cray
June 4, 1995
Acknowledgements

- **BICB**
  - Prof Neuhauser
  - Michael Olesen
  - Steve Smith
- **Cray Inc Resources**