Beyond the Desktop

- The role of computational architectures in accelerating discovery

- Mohammed Khaleel, Ph.D.
High-performance computing systems

*Beyond the Desktop*

- Traditional (or “mainstream”) supercomputers
  - Science applications
- Multithreaded supercomputers
  - Cybersecurity applications
- Energy Efficiency
- Back to the Desktop
Nowadays, HPC systems are *parallel* computing systems

- Consisting of hundreds of processors (or more)
- Connected by high bandwidth, low-latency networks
  - Collections of PCs connected by Ethernet *are not* HPC systems
- Basic building block is a *node*: server-like computer (a few processor sockets, memory and network interconnect cards, possibly I/O devices).

Nodes are parallel computers on their own: contain usually >= 2 processor sockets with multiple cores per processor

- Looks very similar to what you have on your desktop PC!!

HPC systems have a multiplicity of applications in scientific and engineering areas: physics, chemistry, biology, material design, mechanical design.
HPC Systems (cont.)

- Two basic kinds of HPC systems:
  - Distributed memory systems
  - Shared memory systems

- Distributed memory HPC systems:
  - Typical HPC system, processors only have direct access to local memory on the node.
  - Remote memory on other nodes must be accessed indirectly via a library call.
  - Can scale to tens and hundreds of thousands of processors (Blue Gene/P @ LLNL, Chinook @ EMSL/PNNL)

- Shared memory HPC systems:
  - Processors have direct access to local memory on the node and to remote memory on other nodes.
  - Speed of access may vary
  - More difficult to scale beyond a few thousand processors (Columbia SGI Altix @ NASA)
Outline

- High-performance computing systems
  - Beyond the Desktop
- Traditional (or “mainstream”) supercomputers
  - Science applications
- Multithreaded supercomputers
  - Cybersecurity applications
- Energy Efficiency
- Back to the desktop
Chinook (supercomputer at EMSL/PNNL)

- 2310 node HP cluster
  - Dual quad-core processors per node
  - Total: 18,480 cores

<table>
<thead>
<tr>
<th>Feature</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interconnect</td>
<td>DDR InfiniBand (Voltaire, Mellanox)</td>
</tr>
<tr>
<td>Node</td>
<td>Dual Quad-core AMD Opteron 16 GB memory</td>
</tr>
<tr>
<td>Local Scratch</td>
<td>400 MB/s, 924GB/s aggregate 440 GB per node. 1 PB aggregate</td>
</tr>
<tr>
<td>Global Scratch</td>
<td>30 GB/s 250 TB total</td>
</tr>
<tr>
<td>User /home</td>
<td>1 GB/s 20 TB total</td>
</tr>
</tbody>
</table>
Chinook cluster architecture

- **PNNL Network**
- **Chinook Ethernet Core**
  - 288 port IB Switch
- **Chinook InfiniBand Core**
  - 288 port IB Switch
  - 288 port IB Switch
  - 288 port IB Switch
- **Computational Unit 6**
- **Computational Unit 12**
- **Phase-1**
  - 600 nodes
- **Phase-2**
  - 2310 nodes
- **Central Storage**
  - /home NFS PolyServe
  - /dtemp SFS (Lustre)
  - 20 TB 1GB/s
  - 250 TB 30 GB/s
- **Login & Admin**
- **40 Gbit**

**Networks and Storage:**
- **Central Storage**
  - 20 TB 1GB/s
  - 250 TB 30 GB/s
- **Login & Admin**
  - /home NFS PolyServe
  - /dtemp SFS (Lustre)
- **Central Storage**
  - 20 TB 1GB/s
  - 250 TB 30 GB/s
- **Login & Admin**
  - /home NFS PolyServe
  - /dtemp SFS (Lustre)
Chinook software scalability

ScalaBLAST scalability plot
NWChem on Chinook (log-log plots)

Si$_{75}$O$_{148}$H$_{66}$ with DFT

- 3554 functions
- 2300 electrons

(H$_2$O)$_9$ with MP2

- 828 functions
- 90 electrons

C$_6$H$_{14}$ with CCSD(T)

- 264 functions
- 50 electrons
Outline

- High-performance computing systems
  - Beyond the Desktop
- Traditional (or “mainstream”) supercomputers
  - Science applications
- Multithreaded supercomputers
  - Cybersecurity applications
- Energy Efficiency
- Back to the desktop
Processor Architecture (cont.)


Memory Wall Problem

Processor speed

Memory speed

L1

L2

L3

memory

Multithreaded Processors

- Commodity memory is slow, custom memory is very expensive:
  - What can be done about it?

- Idea: cover latency of memory loads with other (useful) computation
  - OK, how do we do this?

- Use multiple execution contexts on the same processor, switch between them when issuing load operations
  - Execution contexts correspond to threads

- Examples: Cray ThreadStorm processors, Sun Niagara 1 & 2 processors, Intel Hyperthreading
Multithreaded Processors (cont.)

Each thread has its own independent instruction stream (program counter)

Each thread has its own independent register set

Execution Units
Cray XMT multithreaded system

- ThreadStorm processors run at 500 MHz
  - 128 hardware thread contexts, each with its own set of 32 registers
  - **No** data cache
  - 128KB, 4-way associative data buffer on the memory side
  - Extra bits in each 64-bit memory word: full/empty for synchronization
  - Hashed memory at a 64-byte level, i.e. contiguous logical addresses at a 64-byte boundary are mapped to uncontiguous physical locations

- Global shared memory
- Scalable to 8,192 processors
Cray XMT multithreaded system (cont.)

- 4 DIMM Slots
- Redundant VRMs
- L0 RAS Computer
High-Performance String Matching on the Cray XMT

- Fast, scalable string matching is at the base of modern cybersecurity applications
  - Deep packet inspection for malware
- Performance has to be consistent and content independent
  - At the same system should be flexible and programmable
  - Prevent content-based attacks
- Excellent scalability and performance on the XMT

![Diagram of Aho-Corasick String Matching]

![Graph showing performance vs. number of processors]
Outline

- High-performance computing systems
  - Beyond the Desktop
- Traditional (or “mainstream”) supercomputers
  - Science applications
- Multithreaded supercomputers
  - Cybersecurity applications

Energy Efficiency

- Back to the desktop
EPA reports energy used in U.S. for servers and data centers is significant.

- ~ 61 billion kilowatt-hours (kWh) in 2006
- 1.5% of total electricity consumption
- Total electricity cost of about $4.5 billion.
- Similar to the amount of electricity consumed by approximately 5.8 million average U.S. households (or about five percent of the total housing stock).
- Federal servers and data centers alone
  - ~ 6 billion kWh
  - 10% of electricity used for servers and data centers
  - Total electricity cost of about $450 million annually.
Current Power Usage by Chinook, MSCF System at PNNL

- Chinook (160TF peak), has 2310 dual socket quad-core AMD Opteron (2.2GHz) based servers from HP each with 16 GB memory, 365 GB local disk, a DDR Infiniband interconnect, and 297 TB global disk
- Consumes nearly 1.9 MW
  - ~ 1/3 for cooling
  - ~ 2/3 compute power (1.25 MW)
    - 40% of compute power is lost to power delivery (rectifier, UPS, PDU, power supply, voltage regulator)
- Average power efficiency for HPL
  - no losses: 133MFlop/s/W
  - with power delivery losses: 80MFlop/s/W
  - with power- and cooling delivery losses: 52MFlop/s/W

40% of compute power lost in power delivery

Top500 measures here
Multiple concurrent basic 4.5 days weather forecasts for North&Central America

- **Initialization**: 1° Global Forecast System analysis from National Weather Service
- **Decomposition**: 480x480 cartesian grid (15km) with 45 levels
- **Solver**: Horizontal: Explicit High-Order Runge-Kutta; Vertical: Implicit
- **Output**: asynchronous 2.3GB netCDF every 3 model-hours per forecast
QM Computational Chemistry (CP2K)

Multiple concurrent liquid-vapor interface model simulations

- **Initialization**: Standard slab geometry (15x15x71 Å³)
- **Decomposition**: 215 H₂O with single hydroxide ion
- **Solver**: Density Functional Theory with dual basis set (Gaussian & Plane-Wave)
  in conjunction with molecular dynamics and umbrella sampling
- **Output**: synchronous 75MB per 20k 0.5fs model-steps (MD time step)
Device Under Test: NW-Ice

- 192 servers, 2.3 GHz Intel (quad-core) Clovertown, 16 GB DDR2 FBDIMM memory, 160 GB SATA local scratch, DDR2 Infiniband NIC
- Five racks with evaporative cooling at processors
- Two racks air cooled
- Lustre Global File System
  - 34TB mounted
  - 49TB provisioned
Contributors to Power Consumption: Power Distribution

Data Center:
- Power Distribution Units
- Power Supply Units
- Voltage Regulators

Facility:
- Transformers
- Rectifiers
- UPS
- Inverters
Contributors to Power Consumption: Cooling Chain

Data Center:
- Air Handlers
- Closely Coupled Cooling Systems
- HVAC

Machine Plant:
- Pumps
- Chillers
- Cooling Towers
- Economizers
Historically, most technologies that have appeared in high-end supercomputers have eventually migrated to the desktop

- Hardware units for numerical computation
- Superscalar execution
- Parallel processing (we’re observing it right now)

In the future, it is expected that most of the technologies I presented today will eventually migrate back to desktop machines

- High-end interconnects between cores & processors
- Multithreading capabilities

Commercial data centers are already looking for ways to improve their energy management