Large Heterogeneous Systems

Sriram Swaminarayan

Team Leader, Evolving Applications & Architectures Team

Los Alamos National Laboratory

A ‘heterogeneous’ system does computation on two or more *different* types of *computation* cores

What is a Heterogeneous System?

- 2 or more different types of computational cores
- Must be ‘current’
- Only ‘large’ systems considered
  - How large is large?
- Reconfigurable systems, such as FPGAs, not considered
- Systems with different (or configurable) networks not considered

Goal is to understand the usability for science

- How vast?
- How fast?
- How painful?
- How portable?

Roadrunner Open Science Lessons Learned - I: Advanced Architectures Are Tractable

- **Wide variety of applications have been accelerated**
- **A graded approach to acceleration is viable**
  - Evolutionary: 2-4x improvement
  - Revolutionary: 6-9x improvement
- **Getting an application running is easy**
- **Getting performance from it requires work**
  - Identical to experience with GPUs today
- **Success requires computer science experts and subject matter experts working together**

Roadrunner Open Science Lessons Learned - II: Keeping Track Of Your Data Is Key To Performance

- **Data Is Everything**
  - Who owns it? (Host or Accelerator?)
  - Where is it now?
  - Where is it needed next?
  - How much does it cost to send it from now to next?

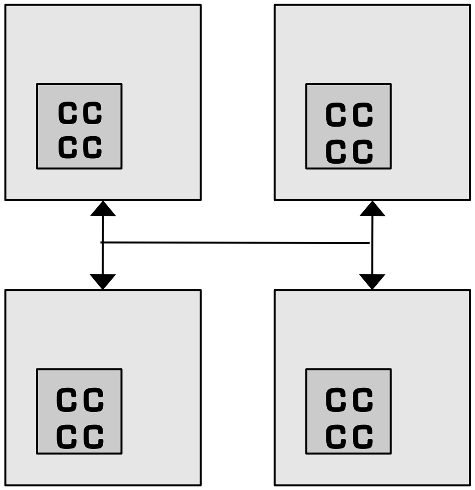
- **Three Primary Data Models Have Emerged**
  - Host Centric: Opteron owns the data
  - Accelerator Centric: Cell (Accelerator) owns data
  - Work Stealing: Dynamic distributed work load

- **Applies directly to almost all heterogeneous clusters**

‘Multicore’ clusters are the norm

- **Cluster made up of identical nodes**
- **Each node has many sockets (~ 4 today)**
- **Each socket hosts a chip with several cores (~ 6 today)**
- **Each core can run many threads (~ 2 today)**
  - AMD / Opterons
  - Intel / Xeons
  - IBM / PowerPC
- **Poster Child: Jaguar 2009**
  - #1 on Top500 list
  - #44 on Green500 list
- **Advantages**
  - All current scientific applications run on these clusters
  - Optimization techniques well understood
  - Compilers are mature
  - Hardware caches insulate scientist from memory hierarchy
- **Disadvantages**
  - Memory bandwidth limited: few codes achieve > 5% of peak performance
  - Power hungry
  - Large number of nodes imply more failure points

Multicore



Vast?	Fast?	Pain?	Portable?
😎	🙄	😎	🙂

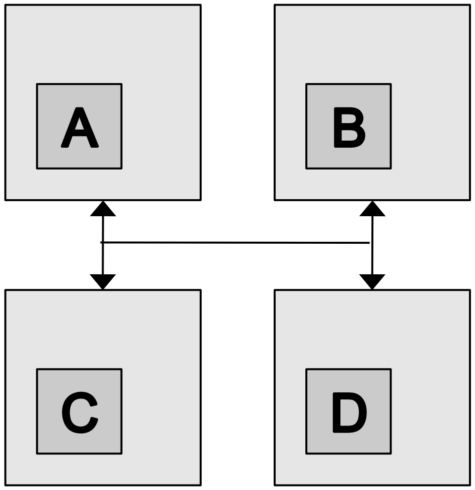
‘Diverse’ clusters are multicore clusters with a mix of different types of nodes

- **Cluster made up of different types of nodes**
- **‘Embarrassingly’ heterogeneous**
- **Each node can have different type of processor**
  - AMD / Opterons
  - Intel / Xeons
  - IBM / PowerPC
- **Poster Child: Jaguar 2008**
  - Cray XT4 & XT5 hooked together
- **Advantages**
  - Almost identical source code across all nodes
  - Inherited from *multicore* clusters

    All current scientific applications run on these clusters
Optimization techniques well understood
Compilers are mature
Hardware caches insulate scientist from memory hierarchy
- **Disadvantages**
  - More than one compiler / binary
  - Partitioning is a challenge
  - Inherited from *multicore* clusters

    Memory bandwidth limited: few codes achieve > 5% of peak performance
Power hogs: most power goes into moving data
Large number of nodes imply more failure points

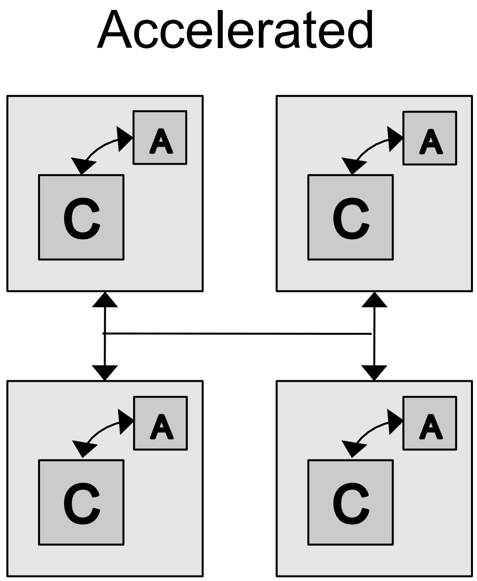
Diverse



	Vast?	Fast?	Pain?	Portable?
Diverse	😎	😕	😊	😊
Multicore	😎	😕	😎	😊

‘Accelerated’ clusters are multicore clusters with computational accelerators on each node

- **Accelerators attached to each compute node**
  - Heterogeneity is off-chip, but on-node
  - Most of compute power resides in accelerator
  - Accelerators not connected to network
- **Each node typically has same type of accelerator**
  - IBM Power XCell 8i
  - Clearspeed
  - GPU (NVIDIA / ATI)
  - MD Grape (special purpose)
- **Poster Child: Roadrunner**
  - Opteron cluster accelerated with Cells
  - #2 on Top500 list
  - #6 on Green500 list
- **Advantages**
  - Power sippers
  - Smaller number of nodes
- **Disadvantages**
  - Cannot easily ‘port’ existing code base
  - Deep memory hierarchy
  - More than one compiler / binary
  - Compilers not mature
  - Partitioning is a challenge
  - Bus bandwidth limited



	Vast?	Fast?	Pain?	Portable?
Accelerated	😬	😊	😢	😢
Multicore	😎	🙄	😎	🙂
Diverse	😎	🙄	🙂	🙂

Each Socket in a ‘Heterogeneous’ cluster contains many different types of cores

- **Heterogeneous chips plugged into each socket**
  - Homogeneous cluster of heterogeneous chips
    i.e. heterogeneity is on-chip
  - Different performance characteristics of each core within each socket
  - Cores can communicate on the network

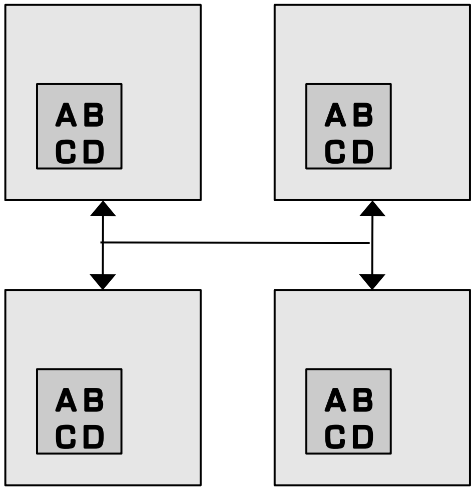
- **All chips are identical: heterogeneity is intra-chip**
  - IBM Power XCell 8i
  - AMD Fusion

- **Poster Child: QPace SFB TR cluster**
  - IBM Power XCell 8is connected with infiniband
  - #110 on Top500 list
  - #1 on Green500 list

- **Advantages**
  - Power misers
  - Smaller number of nodes

- **Disadvantages**
  - Cannot easily ‘port’ existing code base
  - Deep memory hierarchy
  - Memory bandwidth limited
  - More than one compiler / binary
  - Compilers not mature
  - Partitioning is a challenge

Heterogeneous



	Vast?	Fast?	Pain?	Portable?
Heterogeneous	😎	😎	😳	😢
Multicore	😎	😕	😎	🙂
Diverse	😎	😕	🙂	🙂
Accelerated	😬	🙂	😢	😢

‘Manycore’ clusters contain many identical cores per socket

- **Identical chips plugged into each socket**
  - What is ‘many’ cores per chip? (>= 16 maybe?)
  - Identical performance characteristics of each core
  - Cores can communicate on the network

- **All cores are identical**

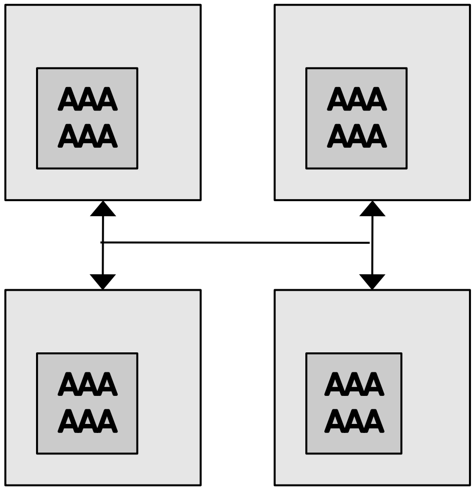
  Typical Manycore chips
  - IBM Blue Gene
  - Intel SCC
  - Sun Niagara
  - Intel Larrabee

- **Poster Child: Dawn** / Sequoia
  - IBM Blue Gene / [P,Q]

- **Advantages**
  - Lower power consumption than multicore clusters
  - Smaller number of nodes
  - Easy to port existing code bases
  - Single compiler / binary

- **Disadvantages**
  - Deep memory hierarchy
  - Memory bandwidth limited

Manycore



AAA
AAA

AAA
AAA

AAA
AAA

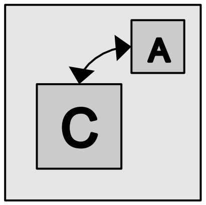
AAA
AAA

	Vast?	Fast?	Pain?	Portable?
Manycore	😎	😎	🙁	🙁
Multicore	😎	😕	😎	🙂
Diverse	😎	😕	🙂	🙂
Accelerated	😬	🙂	😢	😢
Heterogeneous	😎	😎	😐	🙁

UNCLASSIFIED

Accelerated Systems - Current

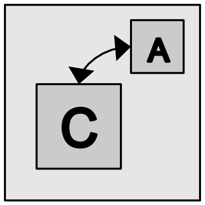
- **Roadrunner – Cell accelerated**
  - Accelerator: Power XCell 8i
  - 96% of compute power in accelerator
  - 1.05 PF (77% of peak)
  - #6 on Green 500 List
  - #2 on Top 500 List
  - Debut 6/08 @ #1
  - Location: LANL
- **Tianhe-1 – GPU accelerated**
  - Accelerator: ATI Radeon HD4870
  - 79% of compute power in accelerator
  - 0.57 PF (47% of peak)
  - #8 on Green500 List
  - #5 on Top500 List
  - Debut 11/09 @ #5
  - Location: NUDT, China



	Vast?	Fast?	Pain?	Portable?
Manycore	😎	😎	🙁	🙁
Multicore	😎	😕	😎	🙂
Diverse	😎	😕	🙂	🙂
Accelerated	😬	🙂	😢	😢
Heterogeneous	😎	😎	😳	🙁

Accelerated Systems - Current

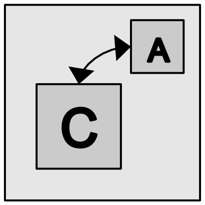
- **Tsubame-1.2 GPU accelerated**
  - Accelerator: NVIDIA Tesla 1070S, ClearSpeed CSX60
  - ??% of compute power in accelerator
  - 0.57 PF (47% of peak)
  - #291 on Green500 List
  - #56 on Top500 List
  - Debut 6/09 @ #41
  - Location: GSIC, Tokyo Institute of Technology



	Vast?	Fast?	Pain?	Portable?
Manycore	😎	😎	🙁	🙁
Multicore	😎	🙄	😎	🙂
Diverse	😎	🙄	🙂	🙂
Accelerated	😬	🙂	😢	😢
Heterogeneous	😎	😎	😐	🙁

Accelerated Systems - Near Future

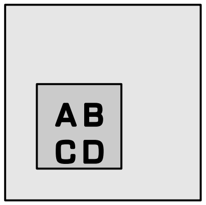
- **Jaguar-2012**
  - Accelerator: NVIDIA Fermi GPU
  - 10-20 PF peak
  - Location: ORNL
- **Tsubame-2.0 2012?**
  - Accelerator: NVIDIA Fermi GPU
  - 3.0 PF peak
  - Location: Tokyo Institute of Technology
- **Keeneland-2012**
  - Accelerator: NVIDIA Fermi GPU
  - 2 PF peak
  - Location: Georgia Tech



	Vast?	Fast?	Pain?	Portable?
Manycore	😎	😎	🙁	🙁
Multicore	😎	😕	😎	🙂
Diverse	😎	😕	🙂	🙂
Accelerated	😬	🙂	😢	😢
Heterogeneous	😎	😎	😐	🙁

Heterogeneous Systems

- **QPace Cluster PowerXCell 8i**
  - Performance: 0.043 PF (77% of peak)
  - #1 on Green500 list
  - #110 on Top500 list
  - Debut 11/09 @ #110
  - Location: Forschungszentrum Juelich, Germany



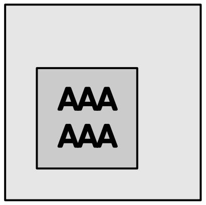
	Vast?	Fast?	Pain?	Portable?
Manycore	😎	😎	🙁	🙁
Multicore	😎	😕	😎	🙂
Diverse	😎	😕	🙂	🙂
Accelerated	😬	🙂	😢	😢
Heterogeneous	😎	😎	😳	🙁

Manycore Systems

- **Dawn: Blue Gene / P**
  - Performance:
  - #22 on Green500 List
  - #11 on Top500 List
  - Debut 6/2009 @ #9
  - Location: LLNL

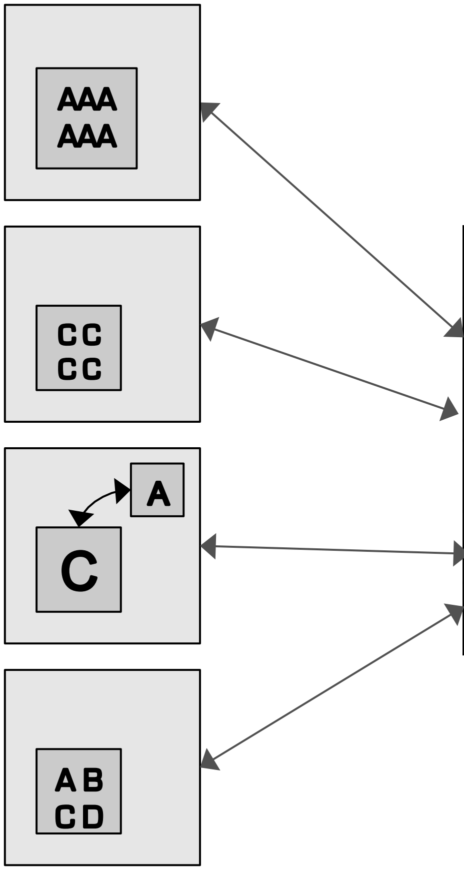
- **Sequoia: Blue Gene/Q**
  - 16 cores? (HPC Wire 2/3/9)
  - 20 PF
  - 2012 delivery
  - Location: LLNL

- **Blue Waters: Power7**
  - 8 cores?
  - 10 PF
  - 2012 delivery
  - Location: UIUC



	Vast?	Fast?	Pain?	Portable?
Manycore	😎	😎	🙁	🙁
Multicore	😎	😕	😎	😊
Diverse	😎	😕	😊	😊
Accelerated	😬	😊	😢	😢
Heterogeneous	😎	😎	😳	🙁

Questions?



	Vast?	Fast?	Pain?	Portable?
Manycore	😎	😎	😟	😟
Multicore	😎	😕	😎	🙂
~~Diverse~~	~~😎~~	~~😕~~	~~🙂~~	~~🙂~~
Accelerated	😬	🙂	😢	😢
Heterogeneous	😎	😎	😳	☹️