# Extending Catamount for Multi-Core Processors

## Cray Users Group

### May 9, 2007

**John Van Dyke, Courtenay Vaughan, Sue Kelly**

jpvandy@sandia.gov, ctvaugh@sandia.gov, smkelly@sandia.gov

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This study was made possible by a special funding by the DOE Office of Science. Part of the testing was conducted at the Oak Ridge National Laboratory.

NNSA National Nuclear Security Administration

Sandia National Laboratories

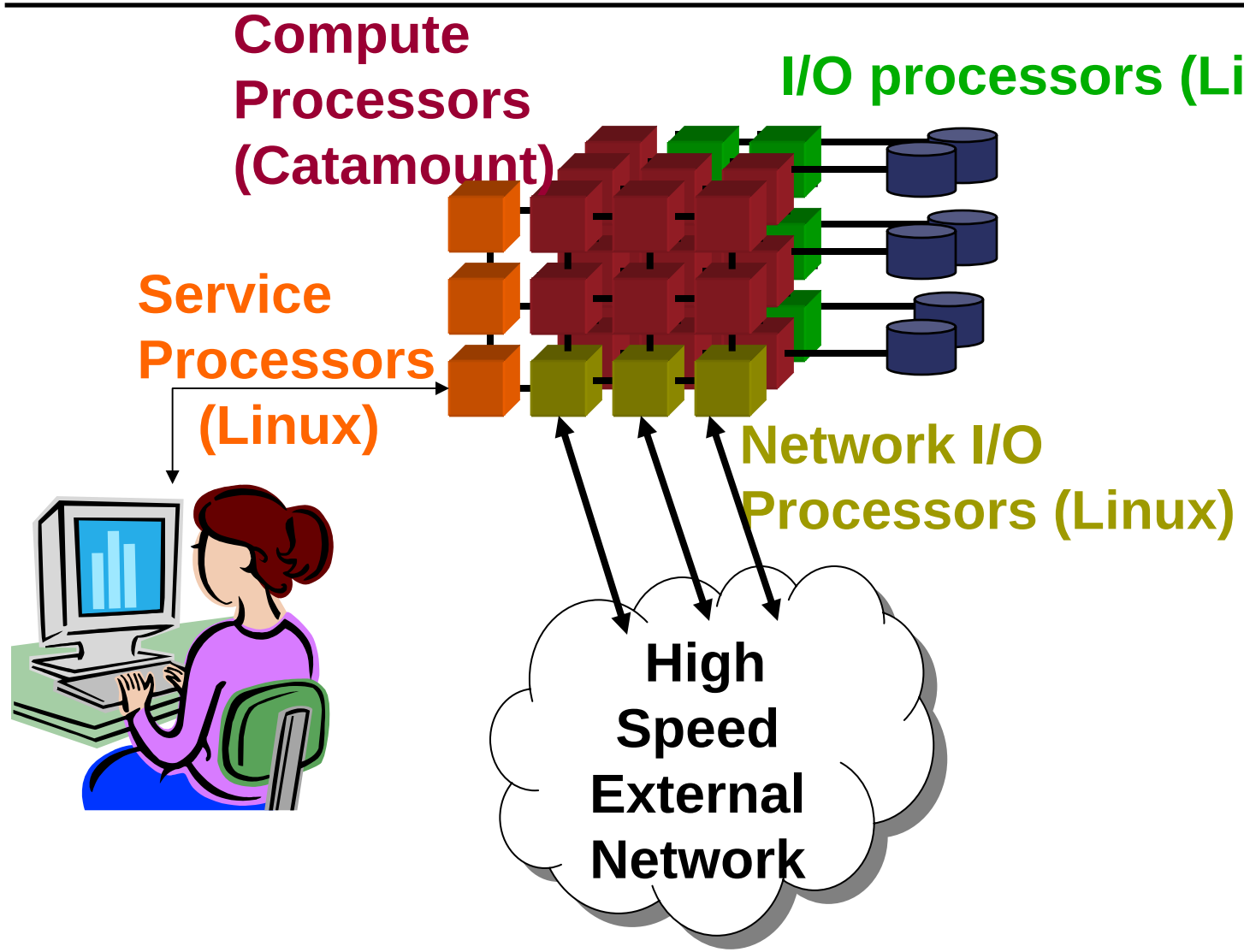
# Catamount for Multi-Core Processors

## Outline

- Overview of Catamount
- Requirements for N-way Catamount
- Design and implementation
- Early dual-core results
- Future

Sandia National Laboratories

# Catamount is designed for an MPP environment with functional partitions




Sandia National Laboratories

# Overview of Catamount

- **LWK – Light Weight Kernel**
- **Catamount OS made up of two pieces**
  - **Quintessential Kernel (QK)**
  - **Process Control Thread (PCT)**
- **Provide functionality necessary to run a scientific calculation.**
- **No disks / no virtual memory / no fork / etc.**
- **Requires high speed network**

Sandia National Laboratories

# Overview of Catamount

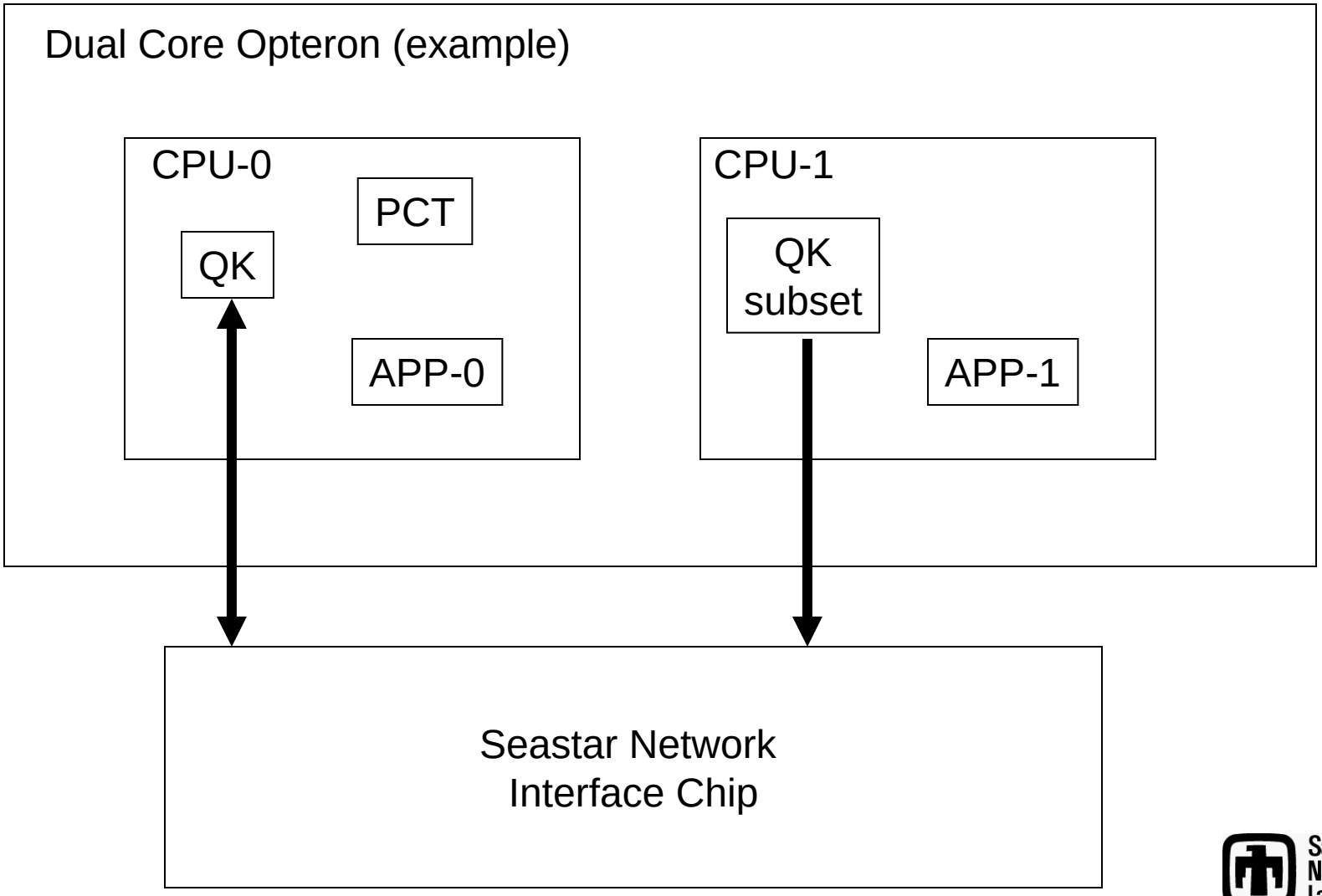
## Virtual Node Mode

**From the Application point of view nearly identical nodes – twice as many -- half the memory**

**From the System point of view, behaves more as master -- slave.**

Sandia National Laboratories

# CPU Responsibility Assignments



# N- way Requirements

- Support 1, 2, or 4 processors/node
  - Desirable: Generalize to N processors/node
- No performance regression between CVN and N-way Catamount on dual core nodes
- MPI and shmem support
- Each core has equal access to NIC for sends
- Support both generic (host-based) and accelerated (NIC-based) portals

# N-way Requirements (2)

- Yod
  - Must be able to specify ppn, processors_per_node, to use
  - Number of virtual nodes does not have to be multiple of ppn.
- Support heterogeneous mode
- Scalable to 100,000 nodes; unlimited virtual nodes
- Minimize OS memory usage; not scale with machine size

# Implications of Requirements

- Common app binary on a node
- Equal division of heap among virtual nodes
- The ppn option is for the job; not the hetero load segment
- # nodes with less than ppn processes on it, is less than ppn
- Process tied to processor
- No OpenMP support
- Share mode not supported

- First six are already true for current CVN

Sandia National Laboratories

# N-way Changes –Design & Implementation

- **Remove PCT arrays dimensioned by # of virtual nodes.**
- **Change binary cpu-0 vs. cpu-1 choices to loops over processors**
- **Adapted the PCT – QK interface**
- **Generalize Process Migration**
- **Yod command line**

  **-sz/-size/-np=#*nodes* [ –ppn=#*processes_per_node*] [ –total-virtual-nodes=#*vn* ]**

- **Generalize QK multi-cpu code**
  - **Separate entries or paths per cpu**
  - **Handling of cpu-id**

Sandia National Laboratories

# N-way Changes –Design & Implementation

**OS memory usage shall not grow with machine size**

- **Remove PCT arrays dimensioned by maximum number of virtual nodes.**
  - **Used in job load**
  - **Borrow application space during load.**
- **One shared table dimensioned by rank of job for the processes on the node.**

# N-way Changes –Design & Implementation

- Change “2” to “N”
- Change binary cpu-0 vs. cpu-1 choices to loops over processors
- Add dimension over cpus to a few structures
- Flag places that are 4-way, not N-way

Sandia National Laboratories

# N-way Changes –Design & Implementation

- **Generalize QK multi-cpu code**
  - **Number of places with separate entries or paths per cpu**
  - **Handling of cpu-id**

- **Flag 4-way code**

# N-way Changes –Design & Implementation

- **Adapted the PCT – QK interface**
  - **Keep track of which “non-cpu-0” process**
  - **Allow passing list of processes/processors**

# N-way Changes –Design & Implementation

## Generalize Process Migration

- Processes start on cpu-0 and “migrate” to another cpu
- Migration is initiated by application (start up library).
- N-way more robust (removes race possibility)
  - Application process requests migration from the PCT
  - PCT requests migration of all processes

- Changes to start-up-library, PCT and QK.

Sandia National Laboratories

# N-way Changes –Design & Implementation

## User API for requesting nodes

- Discontinue use of “-VN” and “-SN”

- Use “-sz/-size/-np” for number of nodes (sockets)
  - This is same number as specified to qsub
- Use “-ppn” for number of processes per node
- Use “-total-virtual-nodes”, if not a multiple of ppn

- Simplest case: all can be omitted and use default

# Test Plan

- Confirm that there are no regressions in N-way from current Catamount Virtual Node (CVN)
  - Verify functionality with test suites
  - Verify performance with applications
- Verify N-way functionality and characterize N-way performance
  - Can use the same tests as above
- Start testing early with baselines from DEV
- Regular testing on Sandia devHarness systems
- Periodic testing on external XT4 systems running DEV

# Current Testing

- John tests very basic functionality on up to 16 dual core nodes as changes are made to the N-way code base. (Hello World, application core-dump, intra-application signaling, etc.)

- Sue verifies functionality with test suites. To date, N-way only tested on 84 single core and 16 dual core nodes.

- Courtenay tests performance using real applications. Tested on Jaguar in April. Jaguar has all dual-core nodes. Results follow.

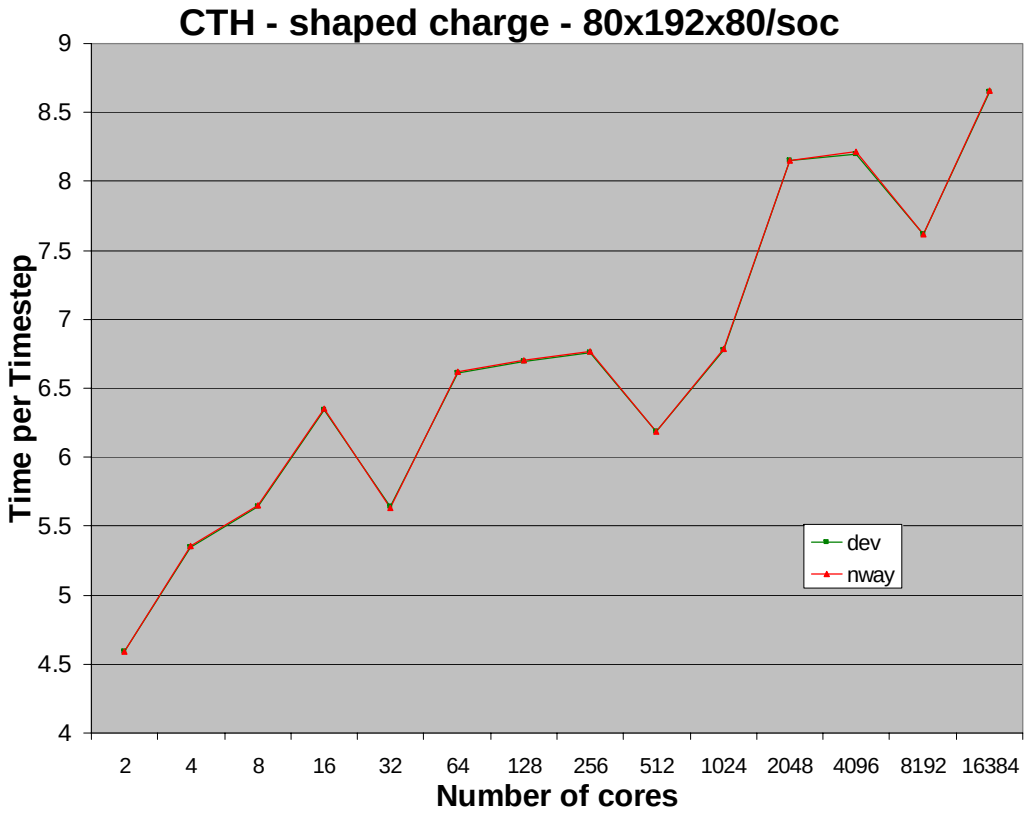
Sandia National Laboratories

# Testing on Jaguar (XT3/XT4) April 23

- Two Applications
  - CTH, a shock hydrodynamics code
  - PARTISN, a neutron transport code
- Problems were scaled with number of processors
- Two series of runs
  - First with CVN
  - Second with N-way
- (Lower on graph is better performance)
- Anomalies all attributed to XT3 – XT4 difference

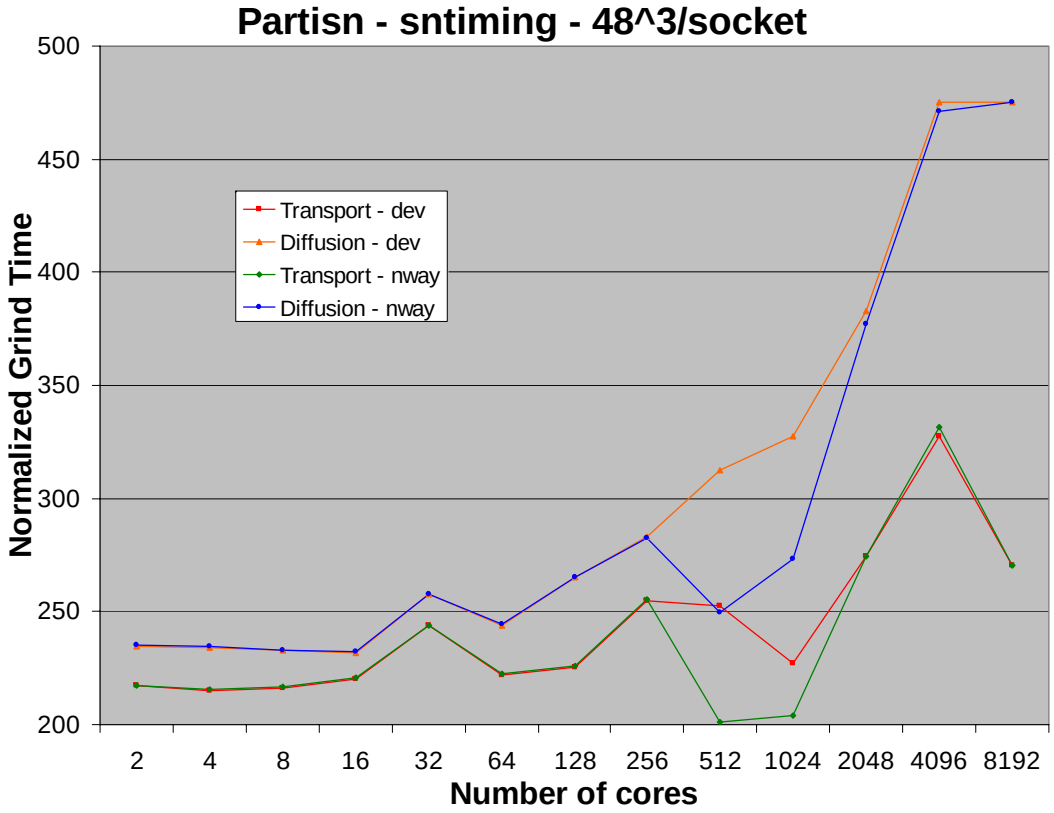
Sandia National Laboratories

# CTH VN Performance




Sandia National Laboratories

# Partisn VN performance




Sandia National Laboratories

# Testing on Jaguar (XT3/XT4) April 23

## Conclusions About April 23rd Tests

**Anomalies attributed to XT3 – XT4 difference.**

**XT4 is faster.**

**No significant difference between CVN and N-way dual-core performance.**

Sandia National Laboratories

# Future

- This is a work in progress
  - Not been on quad core yet
  - To do:  1 gigabyte page support
- Considering SMP node numbering
  - Might relax the heterogeneous: only one ppn value
- Testing, Testing, Testing
  - Quad-core functionality, performance, scaling

Sandia National Laboratories