A Lightweight Kernel Operating System for PetaFLOPS-Era Supercomputers (AKA The Lightweight Kernel Project)

Overview and Current Status

Ron Brightwell

Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy under contract DE-AC04-94AL85000.

Outline

- History
- Project overview
- Current status
- Future directions

Original LWK Project Goals

- Three-year project to design and implement next-generation lightweight kernel for compute nodes of a distributed memory massively parallel system
- Assess the performance and reliability of a lightweight kernel versus a traditional monolithic kernel
- Investigate efficient methods of supporting dynamic operating system services
- Leverage open-source OS projects as much as possible

Original Approach

- Port Cougar LWK to Cplant™ cluster and perform a direct comparison with Linux
  - Performance
  - Scalability
  - Determinism
  - Reliability

Limitations of Original Approach

- Cougar
  - – Not open-source
  - – Export controlled
  - – Not portable
  - – Old
- Cplant™
  - – Alpha is gone
  - – Old

Current Approach

- Short-term
  - Compare Cougar and Linux on ASCI/Red hardware
- Beyond that
  - Figure out how best to leverage Linux or other open-source operating systems to achieve important characteristics of previous LWKs
  - Provide a basis for future OS research activities

Motivation for Linux/Cougar Comparison

- No direct comparison of LWK versus full-service OS since SUNMOS versus OSF1/AD nearly ten years ago
- Much has changed (improved?) since
- A direct comparison between a LWK and Linux is important for providing insight into what is important
- Platform balance is important
- Need real numbers to show people like (Beckman|Minnich|Riesen|Camp)

Sandia National Laboratories

ASCI Red Hardware

- 4640 compute nodes
  - Dual 333 MHz Pentium II Xeons
  - 256 MB RAM
- 400 MB/sec bi-directional network links
- 38x32x2 mesh topology
- Red/Black switchable
- First machine to demonstrate 1+ TFLOPS
- 2.38/3.21 TFLOPS
- Deployed in 1997

ASCI Red Development Systems

- Polaris
  - 8 nodes
  - 200 MHz Pentium Pro
  - Everything else is the same
    - Same memory subsystem
- Nighten
  - 144 nodes
  - Identical hardware as production ASCI Red machine

ASCI Red Compute Node Software

- Puma lightweight kernel
  - Follow-on to Sandia/UNM Operating System (SUNMOS)
  - Developed for 1024-node nCUBE-2 in 1993 by Sandia/UNM
  - Ported to 1800-node Intel Paragon in 1995 by Sandia/UNM
  - Ported to ASCI Red in 1996 by Intel and Sandia
  - Productized as “Cougar” by Intel

ASCI Red Software (cont'd)

- Cougar
  - Space-shared model
  - Exposes all resources to applications
  - Consumes less than 1% of compute node memory
  - Four different execution modes for managing dual processors
  - Portals 2.0
    - High-performance message passing
    - Avoid buffering and memory copies
    - Supports multiple user-level libraries (MPI, Intel N/X, Vertex, etc.)

Cougar Goals

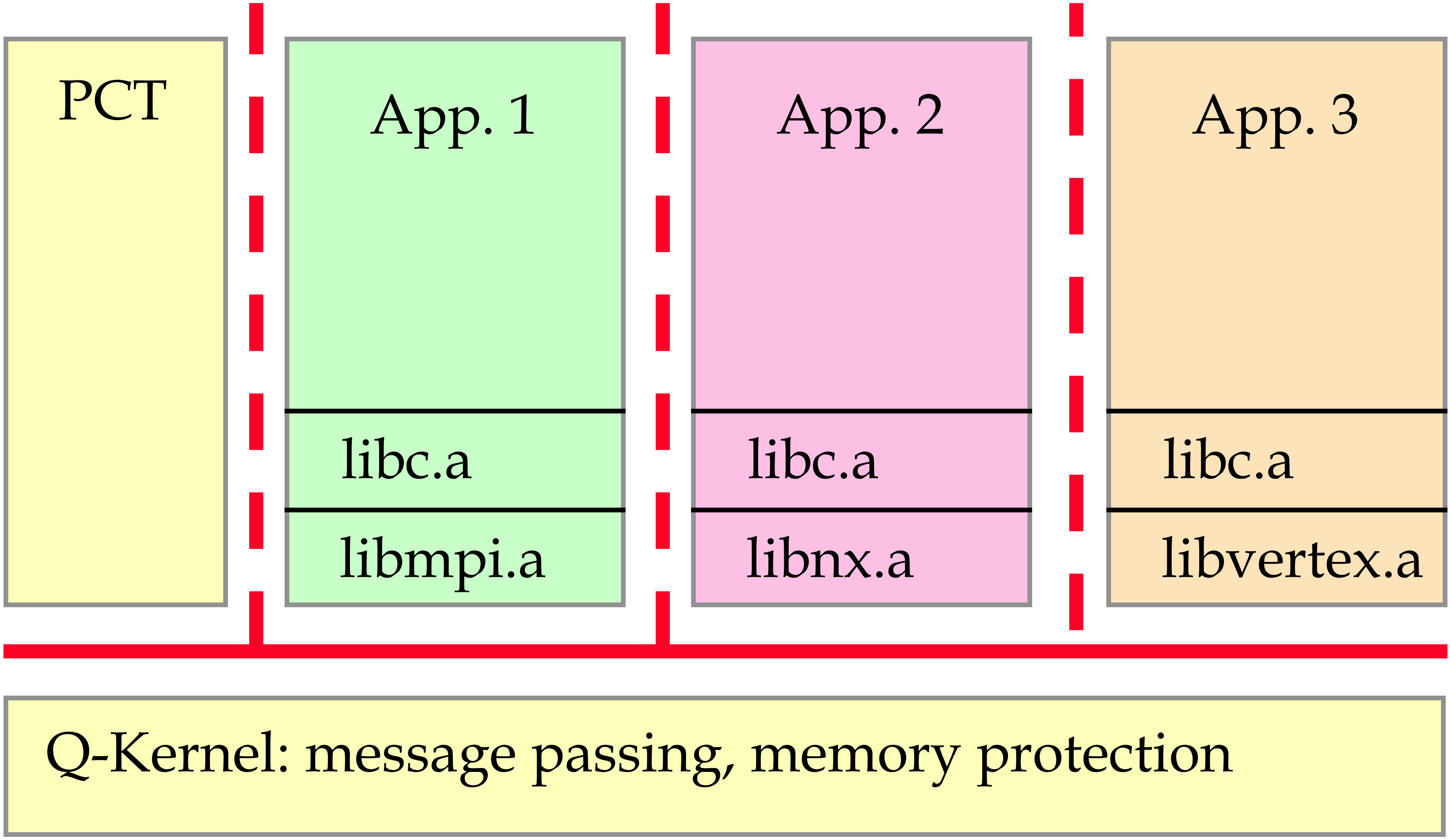
- Targets high performance scientific and engineering applications on tightly coupled distributed memory architectures
- Scalable to tens of thousands of processors
- Fast message passing and execution
- Small memory footprint
- Persistent (fault tolerant)

Cougar Approach

- Separate policy decision from policy enforcement
- Move resource management as close to application as possible
- Protect applications from each other
- Let user processes manage resources
- Get out of the way

Sandia National Laboratories

Cougar General Structure



PCT

App. 1

libc.a

libmpi.a

App. 2

libc.a

libnx.a

App. 3

libc.a

libvertex.a

Q-Kernel: message passing, memory protection

Cougar Quintessential Kernel (QK)

- Policy enforcer
- Initializes hardware
- Handles interrupts and exceptions
- Maintains hardware virtual addressing
- No virtual memory support
- Static size
- Small size
- Non-blocking
- Few, well defined entry points

Cougar Process Control Thread (PCT)

- Runs in user space
- More privileged than user applications
- Policy maker
  - Process loading
  - Process scheduling
  - Virtual address space management
  - Name server
  - Fault handling

Cougar PCT (cont'd)

- Customizable
  - Single-tasking or multi-tasking
  - Round robin or priority scheduling
  - High performance, debugging, or profiling version
- Changes behavior of OS without changing the kernel

Cougar Processor Modes

- Chosen at job launch time
- Heater mode (proc 0)
  - QK/PCT and application process on system CPU
- Message co-processor mode (proc 1)
  - QK/PCT on system CPU
  - Application process on second CPU
- Compute co-processor mode (proc 2)
  - QK/PCT and application process on system CPU
  - Application co-routines on on second CPU
- Virtual node mode (proc 3)
  - QK/PCT and application process on system CPU
  - Second application process on second CPU

Sandia
National
Laboratories

Linux on ASCI Red

- RedHat 7.2 - Linux 2.4.18
- Adapted Linux bootloader and startup code to work with bootmesh protocol
- Service node receives Linux kernel via bootmesh and root filesystem from attached SCSI disk
- Compute nodes mount root filesystem from service node
- Sparse compute node services
  - sshd for remote access
  - Enough libraries for MPI jobs to run

Linux IP Implementation for ASCI Red

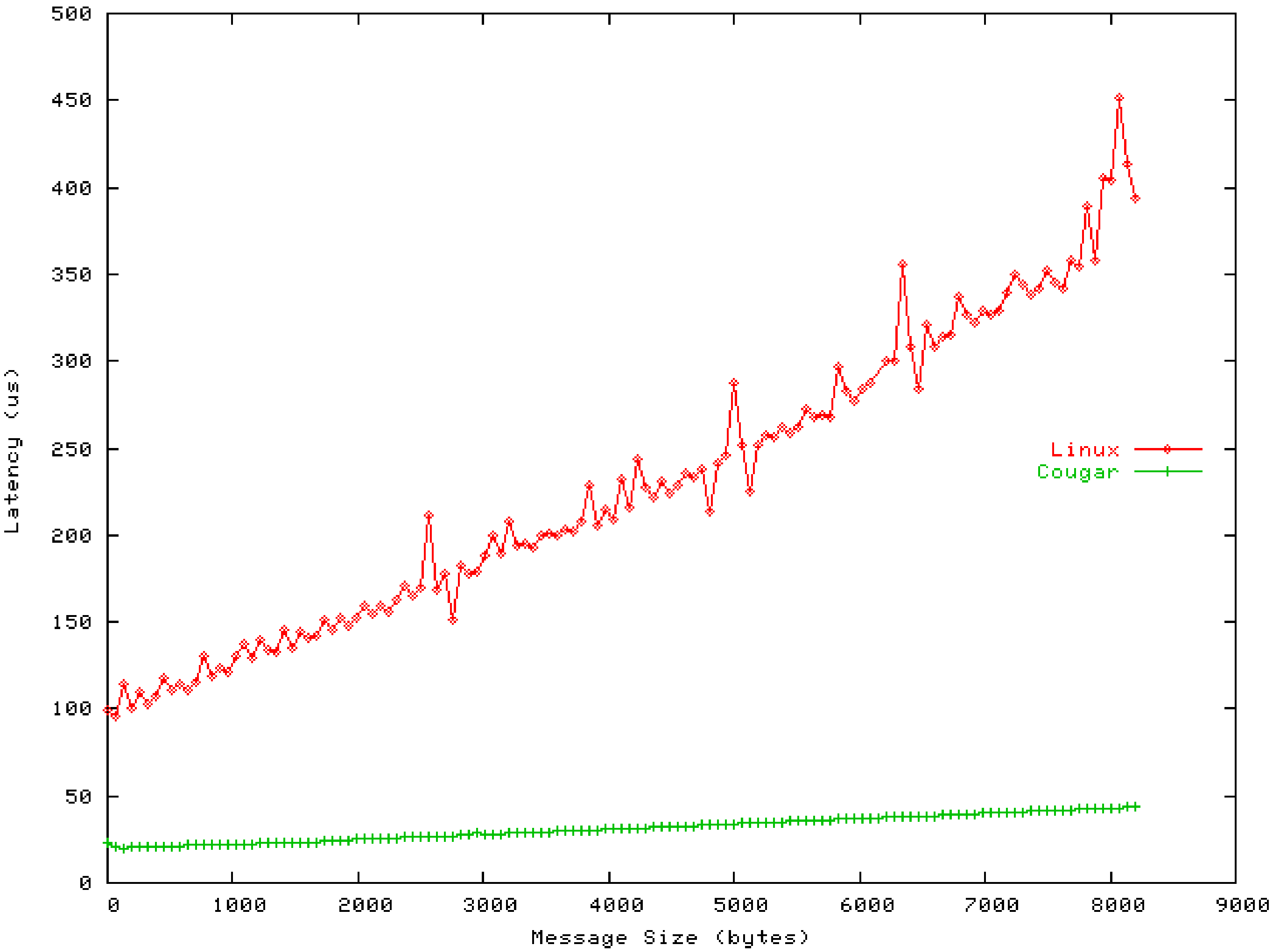
- Implemented a Linux network driver for CNIC
  - Interrupt-driven ring buffer
  - Based on `isa-skeleton.c`
- Varying IP MTU from 4 KB (1 page) to 16 KB (4 pages) showed no noticeable difference in bandwidth
- Bandwidth is CPU limited
  - 45 MB/s for 333 Mhz processors
  - 32 MB/s for 200 MHz processors
- Custom raw device achieved 310 MB/s

Sandia National Laboratories

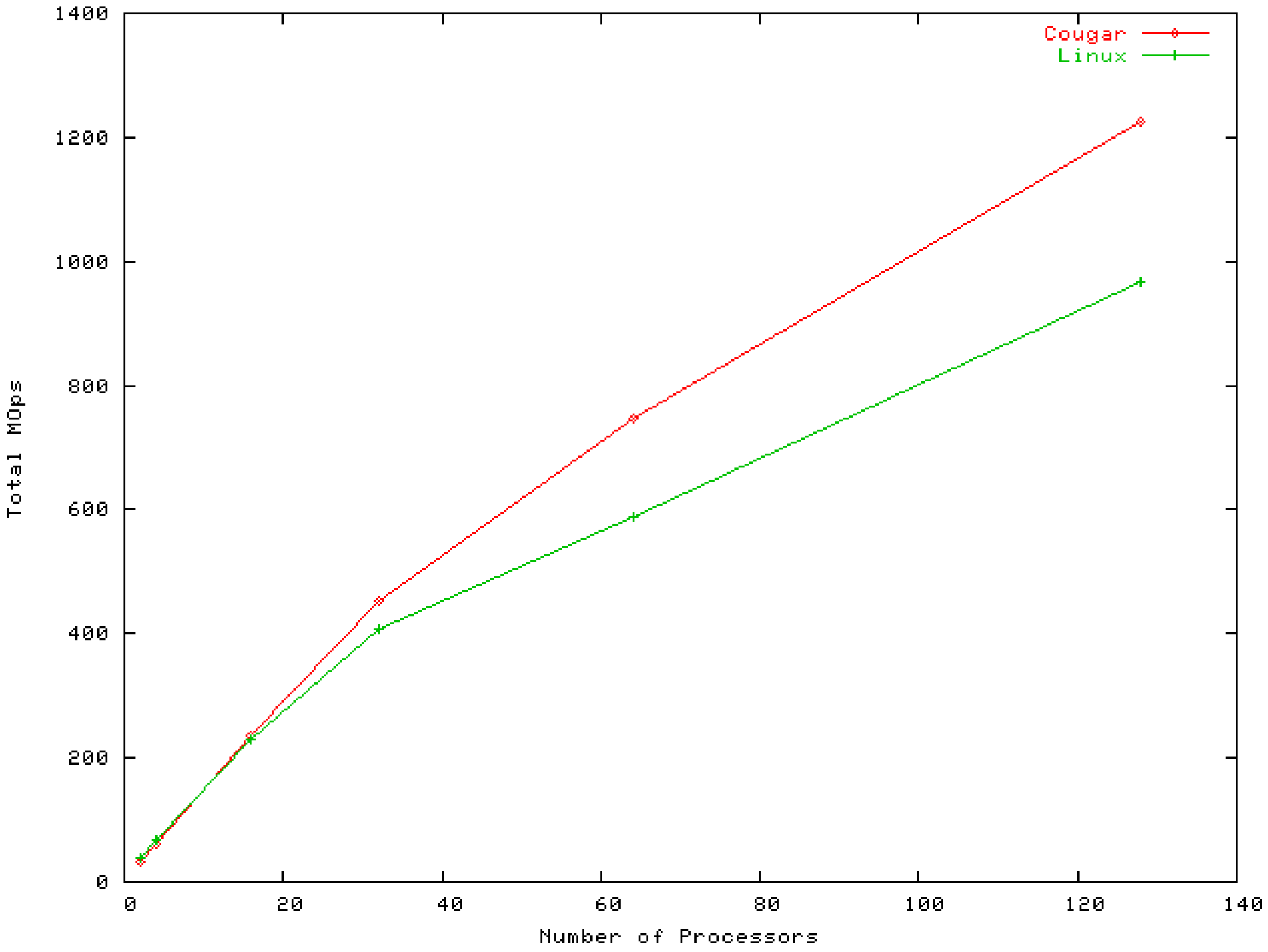
Linux Processor Modes

- Modified CNIC driver to support Cougar processor modes
  - Little difference in performance due to interrupts
- Virtual node mode is default

MPI Ping-Pong Latency

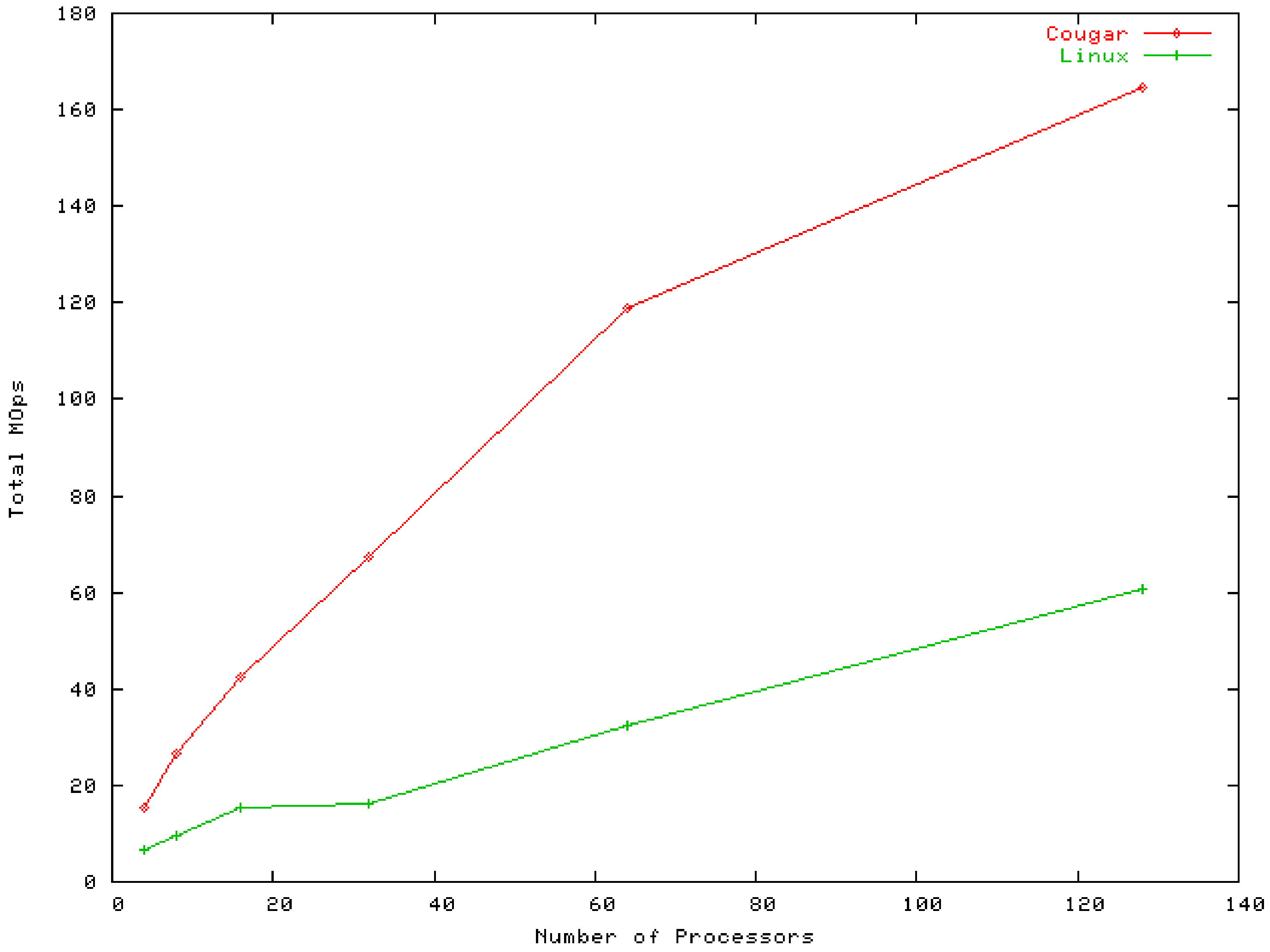


NPB 2.4 - CG



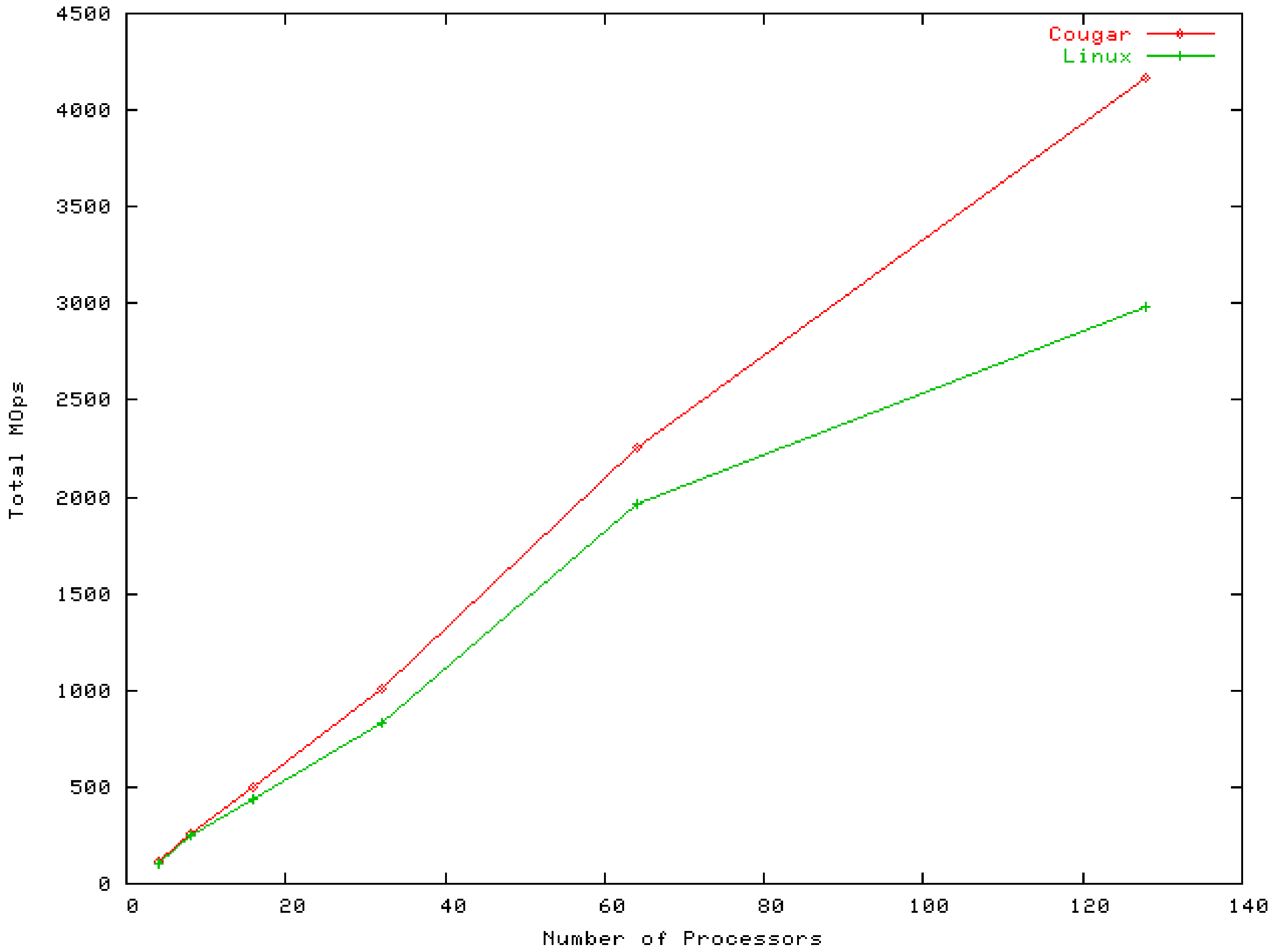
Cougar
Linux
Total MOps
0
200
400
600
800
1000
1200
1400
0
20
40
60
80
100
120
140
Number of Processors

NPB 2.4 - IS



Cougar
Linux
Total MOps
Number of Processors
0
20
40
60
80
100
120
140
160
180
0
20
40
60
80
100
120
140

NPB 2.4 - MG


Cougar
Linux
Total MOps
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100
120
140
Number of Processors

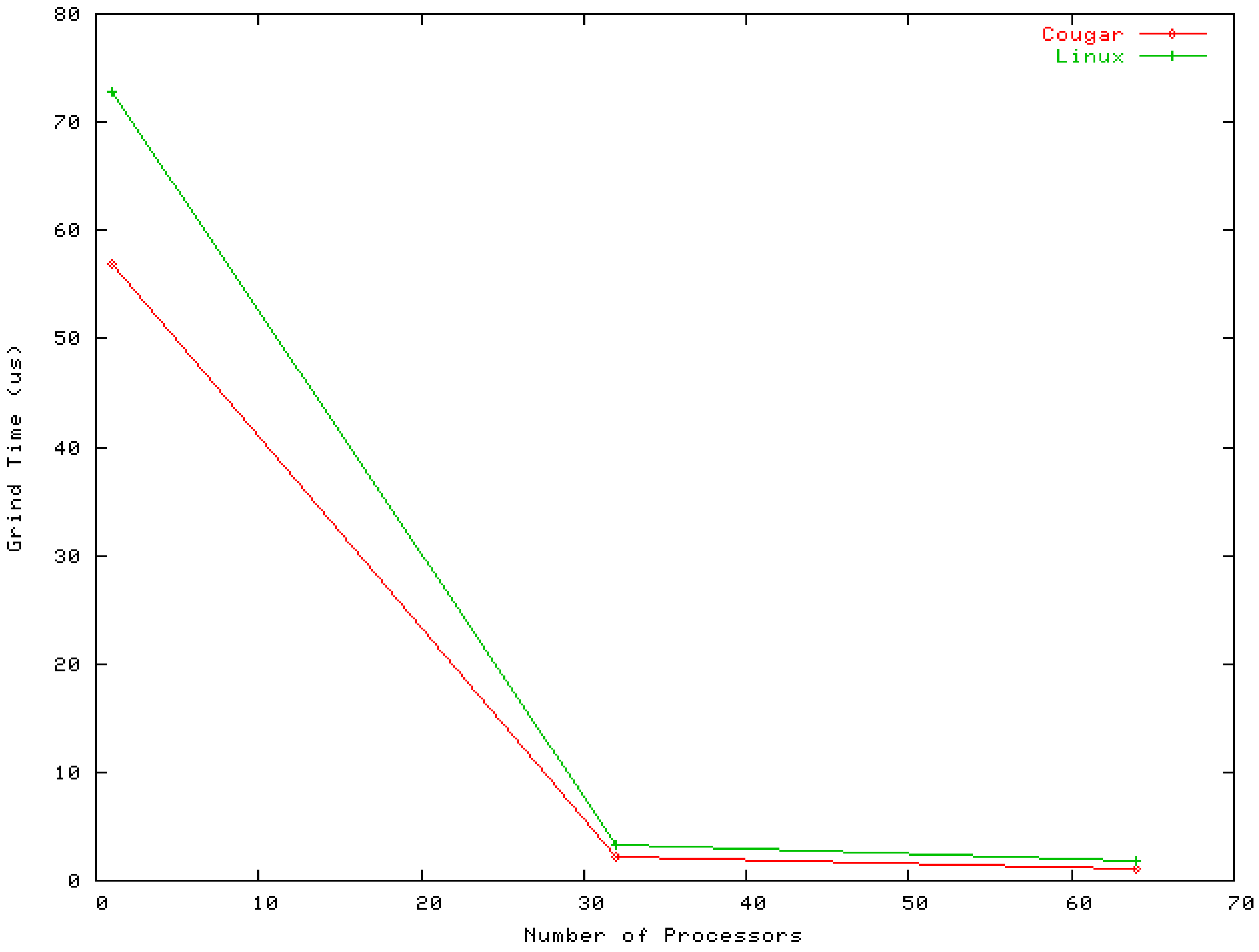
CTH Family of Codes

- **Models complex multi-dimensional, multi-material problems characterized by large deformations and/or strong shocks**
- **Uses two-step, second-order accurate finite-difference Eulerian solution**
- **Material models for equations of state, strength, fracture, porosity, and high explosives**
- **Impact, penetration, perforation, shock compression, high explosive initiation and detonation problems**

CTH Steps

- CTHGEN
  - Problem setup
    - Create computational mesh, insert materials, calculate volume fraction of each material in cells
  - Assign material properties and run-time controls
    - Broadcasting data is main type of message passing
  - Generate initial restart file, one file per node
- CTH
  - Read initial restart file, one file per node
  - Simulate shock wave physics
    - Many nearest-neighbor communications, a few global reductions per time step
  - Write results to restart, history, and viz files
  - Performance measured in grind time
    - Time to compute all calculations on a single cell for a single time step

CTH Performance



Issues

- Compilers and runtime
  - Cougar numbers are from (old) PGI compilers
  - Linux numbers are from (new) Intel compilers
- Determinism
  - No variability in Cougar execution times
    - Even on a loaded machine
  - Significant (>5%) variability in Linux execution times
- Level of effort
  - Maintaining LWK may be equivalent to maintaining a Linux driver

Ongoing Activities

- Completed implementation of Portals 3.2 CNIC driver in Linux
  - 55 µs latency, 296 MB/s
- Currently gathering data for NPB and CTH
  - Need to debug MPI implementation and runtime system
- Linux 2.5
  - Large page support
- Cougar
  - Provide a modern set of compilers/libraries

Conclusions

- Don’t have a real apples-to-apples comparison yet
- Will have a Granny Smith-to-Red Delicious comparison soon

Acknowledgments

- Project team
  - Marcus Epperson, Mike Levenhagen, Rolf Riesen, Keith Underwood, Zhaofang Wen (Sandia)
  - Kurt Ferreira (HP)
  - Trammell Hudson (OS Research, Inc.)
  - Patrick Bridges, Barney Maccabe (UNM)