Quad-core Catamount and R&D in Multi-core Lightweight Kernels

Salishan Conference on High-Speed Computing
Gleneden Beach, Oregon
April 21-24, 2008

Kevin Pedretti
Senior Member of Technical Staff
Scalable System Software, Dept. 1423
ktpedre@sandia.gov
SAND Number: 2008-1725A
Outline

• Introduction

• Quad-core Catamount LWK results

• Open-source LWK

• Research directions

• Conclusion
Going on Four Decades of UNIX

Operating System = Collection of software and APIs
Users care about environment, not implementation details
LWK is about getting details right for scalability
LWK Overview

- POSIX-like environment
- Inverted resource management
- Very low noise OS noise/jitter
- Straight-forward network stack (e.g., no pinning)
- Simplicity leads to reliability
Lightweight Kernel Timeline

1990 – Sandia/UNM OS (SUNMOS), nCube-2
1991 – Linux 0.02
1993 – SUNMOS ported to Intel Paragon (1800 nodes)
1993 – SUNMOS experience used to design Puma
    First implementation of Portals communication architecture
1994 – Linux 1.0
1995 – Puma ported to ASCI Red (4700 nodes)
    Renamed Cougar, productized by Intel
1997 – Stripped down Linux used on Cplant (2000 nodes)
    Difficult to port Puma to COTS Alpha server
    Included Portals API
2002 – Cougar ported to ASC Red Storm (13000 nodes)
    Renamed Catamount, productized by Cray
    Host and NIC-based Portals implementations
2004 – IBM develops LWK (CNK) for BG/L/P (106000 nodes)
2005 – IBM & ETI develop LWK (C64) for Cyclops64 (160 cores/die)

Nov 2007 Top500
Top 10 System
Compute Processors: 82% run a LWK
Challenge: Exponentially Increasing Parallelism

- 900 TF
- 75K cores
- 12 GF/core (89% per year)
- 33% per year

See Key for Units

- TFLOPS
- Cores Past
- Cores Future

2019
- 1 EF
- 1.7M cores (green)
  - 588 GF/core
- 28M cores (blue)
  - 35 GF/core

72% per year
We Know OS Noise Matters

- Impact of noise increases with scale (basic probability)
- Multi-core increases load on OS
- Idle noise measurements distort reality
  - Not asking OS to do anything
  - Micro-benchmark != real application

Red Storm Noise Injection Experiments

- **Result:**
  - Noise duration is more important than frequency
- **OS should break up work into many small & short pieces**
- **Opposite of current efforts**
  - Linux Dynaticks
- **Cray CNL with 10 Hz timer had to revert back to 250 Hz due to OS noise duration issues**

From Kurt Ferreira’s Masters Thesis
Drivers for LWK Compute Node OS

• Practical advantages
  – Low OS noise
  – Performance – tuned for scalability
  – Determinism – inverted resource management
  – Reliability

• Research advantages
  – Small and simple
  – Freedom to innovate (see “Berkeley View”)
    • Multi-core
    • Virtualization
  – Focused on capability systems

• Can’t separate OS from node-level architecture

Much simpler to create LWK than mainstream OS
Outline

• Introduction

• Quad-core Catamount LWK results

• Open-source LWK

• Research directions

• Conclusion
Quad-core Catamount

• Risk mitigation for ORNL Jaguar System
  – Plan of record: CNL + ALPS
  – Backup plan: Quad-core Catamount

• Funded by DOE Office of Science and ORNL
  – PI: Sue Kelly; John VanDyke, Courtenay Vaughan
  – Project complete, fully functional
  – Will be used for Red Storm quad-core upgrade:
    38400 cores, 284 TFLOPS

• Results discussed:
  – Large-scale dual-core CNL vs. Catamount
  – Small-scale quad-core performance
**Large-scale Dual-core CNL vs. Catamount**

<table>
<thead>
<tr>
<th></th>
<th>CNL 2.0.03+</th>
<th>Catamount 2.0.05+</th>
<th>CNL vs. Catamount</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PGI 6.1.6</td>
<td>PGI 6.1.3</td>
<td>% CNL worse</td>
</tr>
<tr>
<td><strong>GTC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024 XT3 only</td>
<td>595.6</td>
<td>584.0</td>
<td>2.0</td>
</tr>
<tr>
<td>4096 XT3 only</td>
<td>614.6</td>
<td>593.8</td>
<td>3.5</td>
</tr>
<tr>
<td>20000 XT3/XT4</td>
<td>786.5</td>
<td>778.9</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>VH1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024 XT3 only</td>
<td>22.7</td>
<td>20.9</td>
<td>8.6</td>
</tr>
<tr>
<td>4096 XT3 only</td>
<td>137.1</td>
<td>117.4</td>
<td>16.8</td>
</tr>
<tr>
<td>20000 XT3/XT4</td>
<td>1186.0</td>
<td>981.7</td>
<td>20.8</td>
</tr>
<tr>
<td><strong>POP</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4800 XT3 only</td>
<td>90.6</td>
<td>77.6</td>
<td>16.8</td>
</tr>
<tr>
<td>20000 XT3/XT4</td>
<td>98.8</td>
<td>75.2</td>
<td>31.4</td>
</tr>
</tbody>
</table>

**Testing performed June 16-17, 2007 at ORNL**
- Apps important to ORNL
- Time ran out before LSMS and S3D problems diagnosed
- Catamount apps did not link with IOBUF library
## Small-scale Quad-core CNL vs. Catamount

<table>
<thead>
<tr>
<th>Application</th>
<th># MPI Ranks</th>
<th>Cores per Node</th>
<th>CNL (time units, lower better)</th>
<th>Catamount (time units, lower better)</th>
<th>(CNL/Catamount - 1) * 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTC</td>
<td>16</td>
<td>4</td>
<td>664.9</td>
<td>670.6</td>
<td>-0.8</td>
</tr>
<tr>
<td>S3D</td>
<td>16</td>
<td>4</td>
<td>1949.1</td>
<td>1948.9</td>
<td>0.0</td>
</tr>
<tr>
<td>POP</td>
<td>16</td>
<td>4</td>
<td>153.8</td>
<td>151.9</td>
<td>1.3</td>
</tr>
<tr>
<td>LSMS</td>
<td>16</td>
<td>4</td>
<td>290.1</td>
<td>276.8</td>
<td>4.8</td>
</tr>
<tr>
<td>SPPM</td>
<td>16</td>
<td>4</td>
<td>847.8</td>
<td>845.0</td>
<td>0.3</td>
</tr>
<tr>
<td>UMT</td>
<td>16</td>
<td>4</td>
<td>8.4</td>
<td>7.9</td>
<td>6.4</td>
</tr>
<tr>
<td>PRONTO</td>
<td>16</td>
<td>4</td>
<td>241.5</td>
<td>222.0</td>
<td>8.8</td>
</tr>
<tr>
<td>SAGE</td>
<td>16</td>
<td>4</td>
<td>267.8</td>
<td>234.9</td>
<td>14.0</td>
</tr>
<tr>
<td>CTH</td>
<td>16</td>
<td>4</td>
<td>15.1</td>
<td>13.0</td>
<td>16.6</td>
</tr>
<tr>
<td>PARTISN</td>
<td>16</td>
<td>4</td>
<td>43.2</td>
<td>35.7</td>
<td>21.0</td>
</tr>
</tbody>
</table>

**Disclaimer:** Some test problems were small

**Testing performed April, 2008 at Sandia**
- Four nodes, 2.2 GHz quad-core, rev. B2
- UNICOS 2.0.44
- 4 KB pages CNL, 2 MB Catamount
- VH1 wouldn’t run on CNL
Catamount Quad-core Cores Effectively Used

<table>
<thead>
<tr>
<th>Application</th>
<th>Utilization of each Core</th>
<th>Cores Effectively Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARTISN</td>
<td>40%</td>
<td>1.60</td>
</tr>
<tr>
<td>CTH</td>
<td>71%</td>
<td>2.84</td>
</tr>
<tr>
<td>SAGE</td>
<td>74%</td>
<td>2.95</td>
</tr>
<tr>
<td>PRONTO</td>
<td>79%</td>
<td>3.18</td>
</tr>
<tr>
<td>UMT2K</td>
<td>91%</td>
<td>3.62</td>
</tr>
</tbody>
</table>

Disclaimer: UMT2K problem was possibly small, others reasonable

Calculation:
- 4 core runs, either 1 core per node (S) or 4 cores per node (Q)
- Assume S takes 1 hr. and Q takes .85 hours
- Assume S using 100% of core
- Q is effectively using \(0.85 \times 4 = 3.4\) of each core
Quad-Core Catamount
Network Stack Performance

• LWK’s static, contiguous memory layout simplifies network stack
  – No pinning/unpinning overhead
  – Send address/length to SeaStar NIC

Host-based Network Stack (Generic Portals)
Testing Performed April 2008 at Sandia, UNICOS 2.0.44
TLB Gets in Way of Algorithm Research

Expected Behavior Due to TLB

Unpredicted

Dashed Line = Small pages

Solid Line = Large pages (Dual-core Opteron)

Open Shapes = Existing Logarithmic Algorithm (Gibson/Bruck)

Solid Shapes = New Constant-Time Algorithm (Slepoy, Thompson, Plimpton)

TLB misses increased with large pages, but time to service miss decreased dramatically (10x). Page table fits in L1! (vs. 2MB per GB with small pages)
Outline

• Introduction

• Quad-core Catamount LWK results

• Open-source LWK

• Research directions

• Conclusion
Project Kitten

• Creating modern open-source LWK platform
  – Multi-core becoming MPP on a chip, requires innovation
  – Leverage hardware virtualization for flexibility
• Retain scalability and determinism of Catamount
• Better match user and vendor expectations

<table>
<thead>
<tr>
<th>COTS</th>
<th>Lightweight Linux</th>
<th>“Rightweight”</th>
<th>LWK</th>
<th>Exotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linux, Solaris, AIX, Windows</td>
<td>Cplant, Cray CNL, Bproc/XCPU</td>
<td>Plan9, K42, L4</td>
<td>Catamount, BGL/CNK, BGP/CNK</td>
<td>Cyclops</td>
</tr>
</tbody>
</table>
Leverage Linux and Open Source

• Repurpose basic functionality from Linux Kernel
  – Hardware bootstrap
  – Basic OS kernel primitives

• Innovate in key areas
  – Memory management (Catamount-like)
  – Network stack
  – Fully tick-less operation, but short duration OS work

• Aim for drop-in replacement for CNL

• Open platform more attractive to collaborators
  – Northwestern and UNM adding their V3VEE lightweight hypervisor to Kitten (NSF funded)
  – Potential for wider impact
Major changes:
- QK includes hypervisor functionality
- QK provides Linux ABI interface, relay to PCT
- PCT provides function shipping, rather than special libc.a
Status

- X86-64 support
- Linux ABI
  - Basic system calls
  - Initial user-stack setup
  - Thread Local Storage (TLS)
  - Virtual system calls
- Boots on Red Storm
  - Drop-in CNL replacement
  - Console I/O
  - Portals network stack
- Initrd treated as PCT (ELF image)
- Runs STREAM compiled with standard Linux toolchain
- DOE approved for open source release (GPL)
make bzImage
make isoimage
kvm -cdrom image.iso
Outline

• Introduction

• Quad-core Catamount LWK results

• Open-source LWK

• Research directions

• Conclusion
SMARTMAP: Simple Mapping of Address Region Tables for Multi-core Aware Programming

Ron Brightwell, Trammell Hudson, Kevin Pedretti

- Leverages LWK memory management model
- Allows all of the processes on a multi-core processor to access each others’ memory directly
  - User-space to user-space
  - No serialization through the OS
  - Access to remote address by flipping a bit
- Each process still has a separate virtual address space
- Allows MPI to minimize memory-to-memory copies on node
  - No copying for non-contiguous MPI datatypes
  - More efficient collective operations
    - Reductions can operate directly on user buffer
Complexity of a Lightweight OS

LWK Code

static void initialize_shared_memory( void )
{
    extern page_table_t *pml4_table_cpu[];
    int cpu;
    for ( cpu=0; cpu < MAX_NUM_CPUS; cpu++ )
    {
        page_table_t *pml4 = pml4_table_cpu[ cpu ];
        if ( !pml4 )
            continue;
        pcb_t *kpcb = cur_kpcb_ptr[ cpu ];
        if ( !kpcb )
            continue;
        page_table_entry_t dirbase = (
            phys_addr( kpcb->kpcb_dirbase )
            | PDE_P
            | PDE_W
            | PDE_U
        );
        int other;
        for ( other=0; other < MAX_NUM_CPUS; other++ )
        {
            page_table_t *other_pml4 = pml4_table_cpu[ other ];
            if ( !other_pml4 )
                continue;
            other_pml4[ cpu+1 ] = dirbase;
        }
    }
}

User Code

static inline void *
remote_address( unsigned core,
    volatile void * vaddr)
{
    uintptr_t addr = (uintptr_t) vaddr;
    addr |= ((uintptr_t) (core+1)) <= 39;
    return (void*) addr;
}
PingPong Latency

- 2.2 GHz Quad-core AMD Opteron
- Catamount N-Way (CNW) 2.0.41
- PGI 7.1.4
- GNU 3.3.3
- Open MPI subversion head
Future Work

- Lots of MPI work
- Expose node/network topology through MPI communicators
  - MPI_COMM_NODE
  - MPI_COMM_NETWORK
- Explore ways for applications to use directly
  - Compiler (BEC)?
  - Libraries (LibSM)?
Mitigating DRAM Bank Conflicts

128KB +/- 16KB Spacing Results In DRAM Conflicts

STREAM Copy w/ DRAM PAGE_CONFLICT Counts

128KB Spacing Worst Case
Application Power Signatures

- Pallas
- HPCC
- HPL
- Catamount Power Saving IDLE
- Compute Node Linux IDLE
Conclusion

• Sandia focusing on needs of capability systems
• Quad-core Catamount ready for action
  – Risk mitigation for ORNL Jaguar
  – Will be used for Red Storm upgrade: 38400 cores, 284 TFLOPS
• Kitten LWK in development
  – Open source
  – Multi-core and hardware virtualization
• Leveraging LWK for system software research
Acknowledgments

• Quad-core Catamount
  – Office of Science and ORNL
  – Sue Kelly, John VanDyke, Courtenay Vaughan, Jim Tomkins

• Kitten LWK
  – Kurt Ferreira, Trammell Hudson, Sue Kelly, Michael Levenhagen, John VanDyke

• SMARTMAP
  – Ron Brightwell, Trammell Hudson

• DRAM Bank Conflicts
  – Kurt Ferreira, Courtenay Vaughan

• Power Signatures
  – Jim Laros