Recent Trends in Operating Systems and their Applicability to HPC

Arthur Maccabe, Patrick Bridges
University of New Mexico

Ron Brightwell, Rolf Riesen
Sandia National Laboratories

May 11, 2006
Lugano, Switzerland
Variety, variety, variety

Mach \rightarrow \text{microkernels} \rightarrow \text{hypervisors} \rightarrow \text{Xen} \\
L4 \rightarrow \text{microkernels} \rightarrow \text{CNK} \rightarrow \text{lightweight kernels} \\
\text{Catamount}
Microkernels

- Minimal services
- Policy versus mechanism
- Address spaces, control (threads), message passing
- Servers
  - Trampoline
Hypervisors

- Hypervisor virtualizes hardware
- goal is to run multiple OSes
- direct access to hardware is preferred
- Xen (para)virtualizes Processor, MMU, and basic I/O
- Additional I/O virtualization done by Domain 0
Lightweight Operating Systems

- Catamount
- SUNMOS, Puma/Cougar
- Catamount, Portals
- Blue Gene/L
- Compute Node Kernel (CNK)
- I/O Nodes (linux)
Blue Gene/L CNK

- I/O nodes
- run Linux
- have storage resources
- separate I/O network
- Compute nodes
- run lightweight kernel
- high speed, partition-able network
Catamount

- QK – mechanism
  - communication
  - address spaces
- PCT – policy
  - finding servers
- Wrapper lib
  - wrapper for stdio calls
- RPC to I/O node
Linux, the 800 Pound Penguin

- Imagine that you are a “small” computer company in the US
- One customer believes in lightweight OSes
- Another demands Linux
- You can’t afford to support the code bases for two OSes
- What do you do?
- The world is waiting for your answer....
What does Linux do?

- Provides a wide range of services
  - libraries
  - development environment
- work environment
- Works on a wide range of hardware
  - graphics cards
  - I/O buses
  - flaky stuff.....
- Hourglass design
What does Linux do in HPC?

- Don’t really have that many devices
- No disks
- None of it is flaky :)
- Must be the services
  - Probably not mail, emacs, or the terminal emulator....
  - “Real men read their mail on a Paragon”
Lightweight Linux?

I’m busy planning to rule the world!

Well, good luck with that.
Running Linux on BG/L

• Seems like a “no-brainer”
  • some people will tell you that BG/L already runs Linux....
• It’s not.....
  • “exec” is reasonable, but what does “fork” mean?
  • what is the right tradeoff for resources allocated to Linux?
  • Is that really Linux on the I/O nodes?
Linux on Catamount

- Basic idea
  - QK == Xen
  - PCT == Dom 0
- QK virtualization
  - PCT builds address spaces
  - PCT can run contexts
  - Portals for network
- Use XenoLinux
  - emulate Xen hypercalls
  - no mod of XenoLinux
# Xen Hypercalls

<table>
<thead>
<tr>
<th>Hypercall</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>set_callbacks</td>
<td>normal and “failsafe” handlers</td>
</tr>
<tr>
<td>sched_op_new</td>
<td>yield, block, shutdown, poll</td>
</tr>
<tr>
<td>mmu_update</td>
<td>update page table entries</td>
</tr>
<tr>
<td>stack_switch</td>
<td>change the stack</td>
</tr>
<tr>
<td>fpu_taskswitch</td>
<td>next use of FPU faults</td>
</tr>
<tr>
<td>memory_op</td>
<td>increase/decrease memory allocation</td>
</tr>
<tr>
<td>event_channel_op</td>
<td>inter-domain event-channel mgmt</td>
</tr>
<tr>
<td>physdev_op</td>
<td>BIOS Replacement</td>
</tr>
</tbody>
</table>
Catamount Mechanisms

Portals + Build Address Space, Run Context
Interrupts: Quantum Network

PCT (Process Control Thread)

illegal instr, addr quit quantum
QK

Hardware

MB
Wrap Lib
Application 1
Wrap Lib
Application 2

Portals + Quit Quantum, Illegal instruction, Illegal address
A more realistic picture

- Start with XenoLinux
- minimize modifications
- build a wedge to provide QK interface
- wedge could support page table construction
- Extend PCT and QK to support XenoLinux
- minimize impact on Catamount applications
- minimize changes to QK
Never forget that the real goal is to run a single application per node (multiple processes, multiple threads)
Why Linux on Catamount?

- Linux is **not** free
  - Initial port and optimization
  - Linux evolves and requires updates
  - Does “lightweight” Linux exist?
- Catamount currently works and scales
  - not clear that Linux will scale
  - Catamount doesn’t evolve :) :)”
- Use XenoLinux on Catamount
  - XenoLinux will evolve: evolve wedge, then PCT; QK only when necessary
- Minimal number of supported code bases
FAST-OS

Forum to Address Scalable Technology for runtime and Operating Systems
<table>
<thead>
<tr>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colony</td>
</tr>
<tr>
<td>Virtualization on minimal Linux with SSI services</td>
</tr>
<tr>
<td>Config</td>
</tr>
<tr>
<td>Combine micro services to build app specific OS</td>
</tr>
<tr>
<td>DAiSES</td>
</tr>
<tr>
<td>Adaptation of OS based on Kperfmon &amp; Kerninst</td>
</tr>
<tr>
<td>K42</td>
</tr>
<tr>
<td>Enhance applicability of K42 for HEC OS research</td>
</tr>
<tr>
<td>MOLAR</td>
</tr>
<tr>
<td>Modules to config and adapt Linux + RAS &amp; fSM</td>
</tr>
<tr>
<td>Peta-Scale SSI</td>
</tr>
<tr>
<td>Intersection of big (SMP) and small (node) kernels</td>
</tr>
<tr>
<td>Right-Weight</td>
</tr>
<tr>
<td>Build application specific Linux/Plan 9 kernels</td>
</tr>
<tr>
<td>Scalable FT</td>
</tr>
<tr>
<td>Implicit, explicit, incremental checkpointing &amp; resilience</td>
</tr>
<tr>
<td>SmartApps</td>
</tr>
<tr>
<td>Vertical integration between SmartApps and K42</td>
</tr>
<tr>
<td>ZeptoOS</td>
</tr>
<tr>
<td>Ultralight Linux, collective runtime, measure &amp; FT</td>
</tr>
<tr>
<td>FAST-OS</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>Colony</td>
</tr>
<tr>
<td>ConfigOS</td>
</tr>
<tr>
<td>DAiSES</td>
</tr>
<tr>
<td>K42</td>
</tr>
<tr>
<td>MOLAR</td>
</tr>
<tr>
<td>Peta-Scale SSI</td>
</tr>
<tr>
<td>Rightweight</td>
</tr>
<tr>
<td>Scalable FT</td>
</tr>
<tr>
<td>SmartApps</td>
</tr>
<tr>
<td>ZeptoOS</td>
</tr>
</tbody>
</table>

| Linux | ✔️ | ✔️ | ✔️ | ✔️ | ✔️ | ½ | |

- **H**: High
- **M**: Medium

**Fault Handling**

**Common API**

**SSI**

**Collective RT**

**I/O**

**OS Noise**
# Partners

<table>
<thead>
<tr>
<th></th>
<th>Lead</th>
<th>Academic</th>
<th>Industrial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colony</td>
<td>LLNL</td>
<td>UIUC</td>
<td>IBM</td>
</tr>
<tr>
<td>Config</td>
<td>SNL</td>
<td>UNM, Caltech</td>
<td>IBM</td>
</tr>
<tr>
<td>DAiSES</td>
<td>UTEP</td>
<td>Wisconsin</td>
<td>IBM</td>
</tr>
<tr>
<td>K42</td>
<td>LBNL</td>
<td>Toronto, UNM</td>
<td>IBM</td>
</tr>
<tr>
<td>MOLAR</td>
<td>ORNL</td>
<td>LaTech, OSU, NCSU</td>
<td>Cray</td>
</tr>
<tr>
<td>SSI</td>
<td>ORNL</td>
<td>Rice</td>
<td>HP, CFS, SGI, Intel</td>
</tr>
<tr>
<td>Right-Weight</td>
<td>LANL</td>
<td></td>
<td>Bell Labs</td>
</tr>
<tr>
<td>Scalable FT</td>
<td>PNNL</td>
<td>LANL, UIUC</td>
<td>Quadrics, Intel</td>
</tr>
<tr>
<td>SmartApps</td>
<td>Texas A&amp;M</td>
<td>LLNL</td>
<td>IBM</td>
</tr>
<tr>
<td>ZeptoOS</td>
<td>ANL</td>
<td>Oregon</td>
<td></td>
</tr>
</tbody>
</table>
FAST-OS

- PI meeting/workshop (open meeting)
- with USENIX in Boston, May 30 & 31
- http://www.cs.unm.edu/~fastos
- Most recent issue of ACM OSR

“Linux’s cleverness is not in the software, but in the development model”