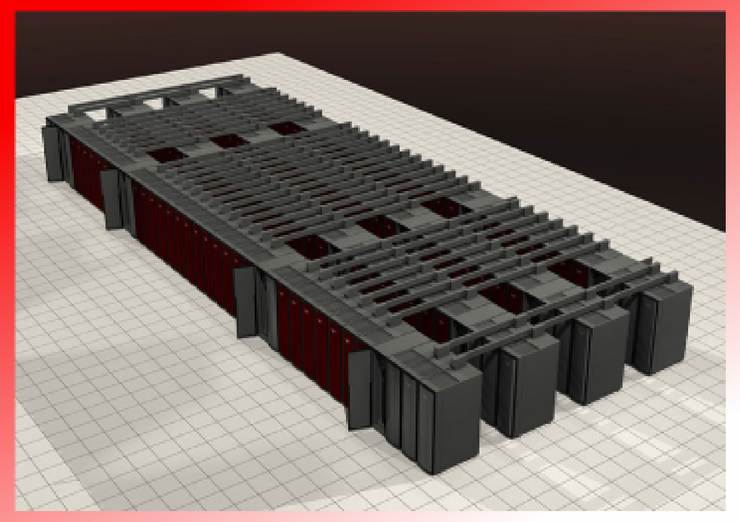
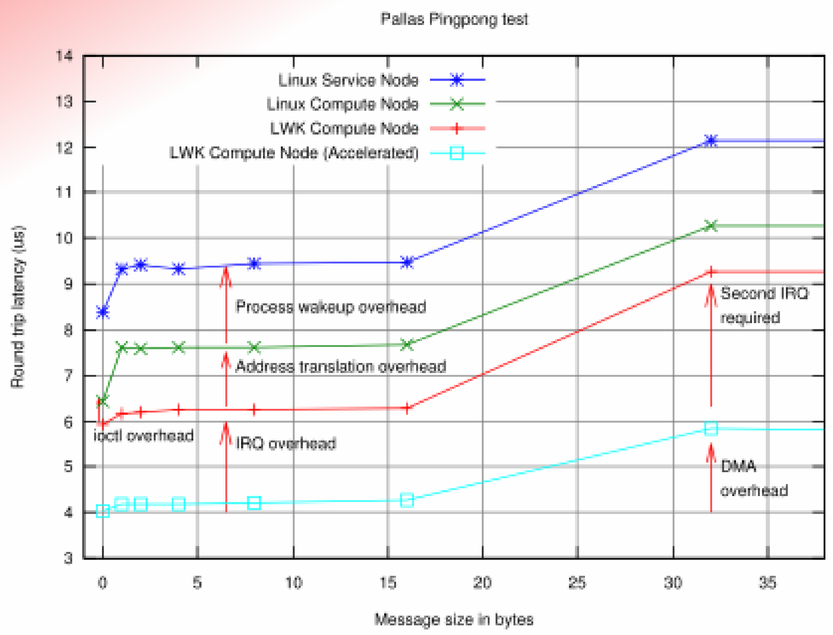


Network Performance Impact of a Lightweight Linux for Cray XT3 Compute Nodes

Trammell Hudson
hudson@osrearch.net

Ron Brightwell
rbbrigh@sandia.gov

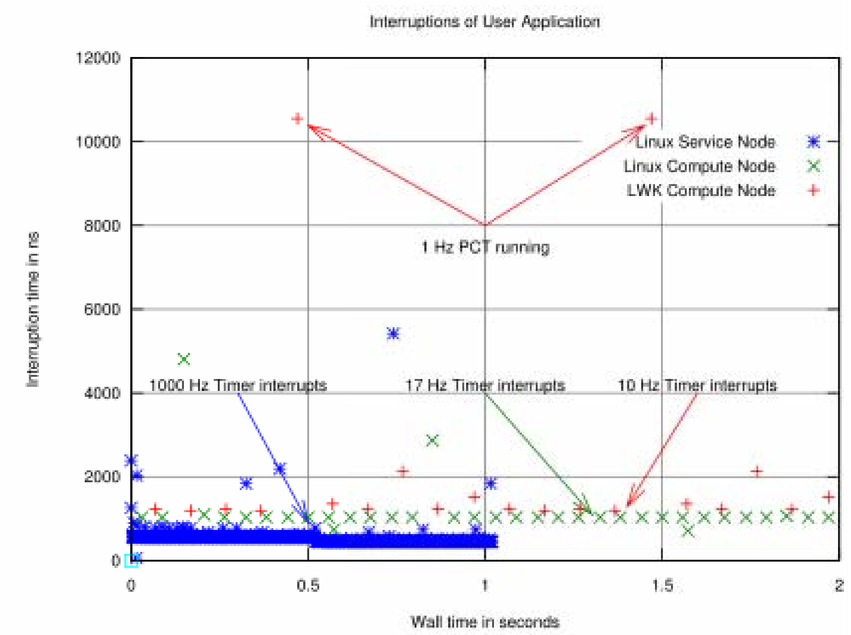


The Catamount Lightweight Kernel (LWK) is a single programmed operating system that allows the user applications to busy wait for events. With a multiprogrammed Linux environment, however, this polling would adversely affect any other processes on the node, so the Portals device driver puts a process to sleep in a select(2) on the file descriptor. This adds 3 µs of latency due to the cost in awakening the process.

Compute node Linux is single programmed, just like Catamount, so we modified the network device driver to allow the application to busy wait. This results in a reduction of 1.8 µs of latency.

All three systems are able to make use of the Portals "Very Small Message" (VSM) optimization that sends 16-byte or smaller messages in the Portals header and only requires a single DMA transfer and a single interrupt to the host. Once messages cross this VSM boundary, an extra interrupt and extra DMA is required to receive it.

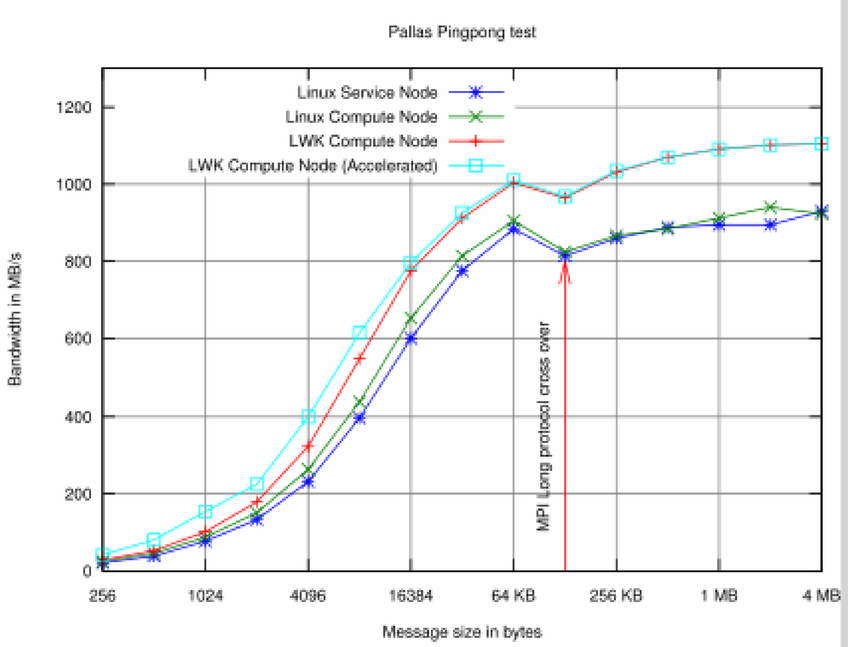
The Catamount LWK also supports an "accelerated" mode for Portals in which the SeaStar network interface processes messages without interrupts. This mode significantly improves short message latency. However, this mode requires a physically contiguous address space, which is not available in Linux.



The Sandia Selfish benchmark is a reimplementation of the ANL Zepto-OS Selfish benchmark. The benchmark spins in a tight loop sampling the CPU timestamp counters until it sees a discontinuity greater than some threshold. This may be caused by an interrupt, a process scheduling event or any other system activity that takes the CPU away from the application.

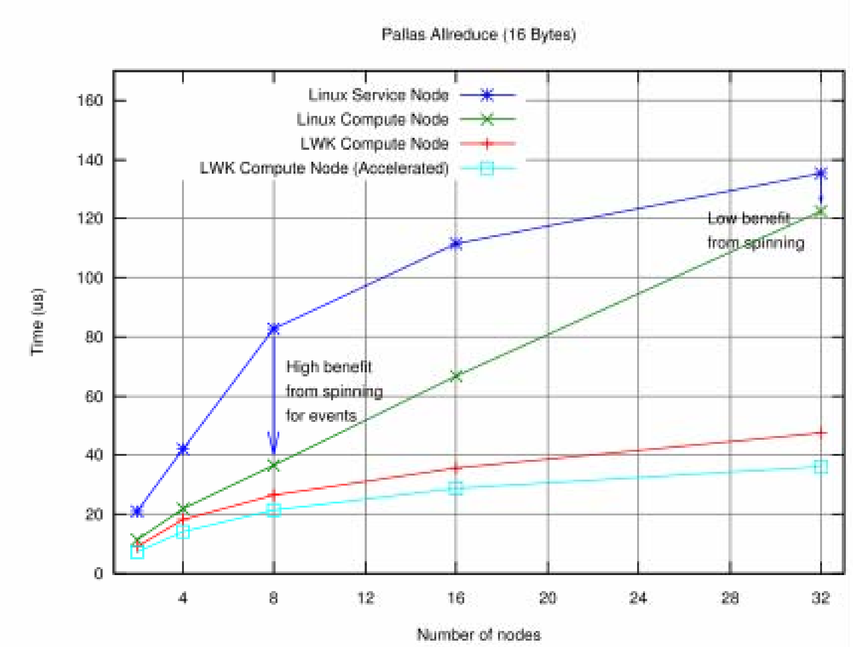
The time series graphically illustrates the overhead of different operating systems and scheduling algorithms. Every interruption may cause TLB entries to be reused, cache lines to be flushed and can contribute to the "rogue OS" effect.

Catamount runs only a 10 Hz timer interrupt and 1 Hz scheduling interval, leading to very minimal overhead. By default Linux uses a 1000 Hz timer interrupt and 60 ms scheduling interval, creating the solid blue line. We modified the kernel to use a 17 Hz timer and increased the maximum quantum to 1 second. This increased the CPU availability of Linux to be slightly better than that of Catamount.



The SeaStar DMA engines are capable of sending or receiving 256 KB in a single transfer, which allows Catamount to use far fewer DMA entries for long message transfers due to its physically contiguous allocation of pages for all application segments. Linux uses small pages and non-contiguous assignment, limiting the maximum DMA transfer to 4 KB and requiring on the order of 64 times more DMA entries than the LWK.

Both systems share a cross over at 128 KB from a short message eager send protocol with MPI buffering of unexpected messages on the receiving end to a long message protocol that ensures that the receive has been posted before the data is transfered.



Even though the single node compute performance of compute node Linux is very close to that of Catamount, it does not scale nearly as well as the LWK. Even at the small scale of 28 nodes, the LWK is far outperforming Linux on collective operations. The Pallas Allreduce benchmark shows that on a small number of nodes compute node Linux is much better than stock Linux and is close to Catamount's performance, mostly due to the decreased latency of event delivery, but as the number of nodes increases, performance of the two Linux kernels converge due to cascading delays in the message handling.

The accelerated interface shows some improvement over the host-based matching, but it is not as significant at this scale as the overhead of the Linux kernel.

It is very interesting that the "rouge OS" effect may be evident at such a small scale, especially since the compute node Linux runtime has slightly higher CPU availability than the Catamount kernel.

National Nuclear Security Administration

Sandia National Laboratories

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.