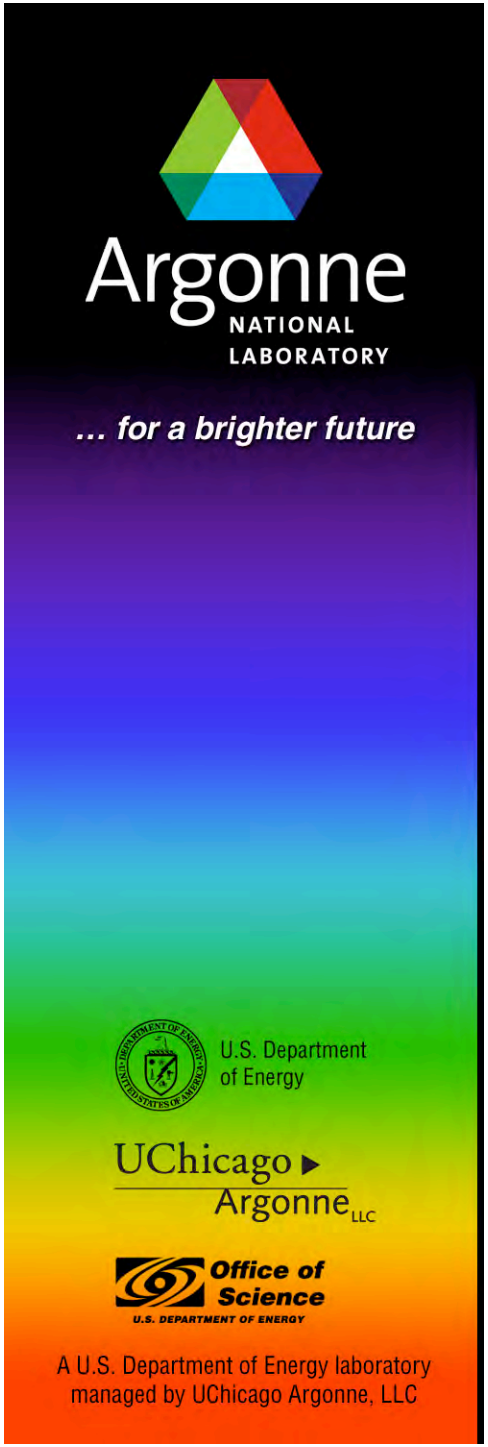


Linux on 100M Processors: Why Not?

Pete Beckman

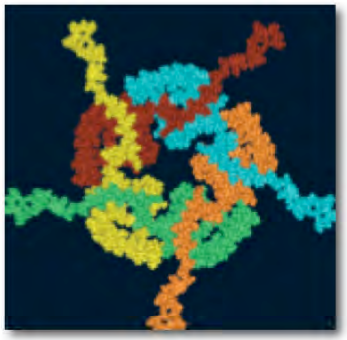
Kazutomo Yoshii

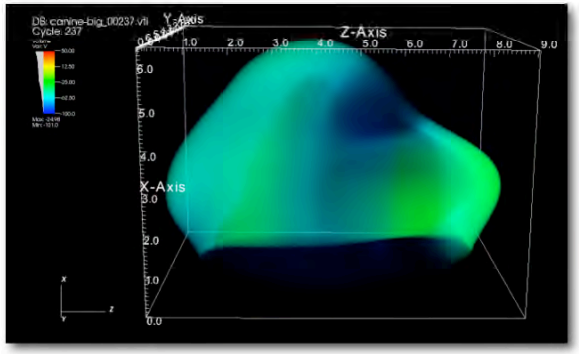
Kamil Iskra

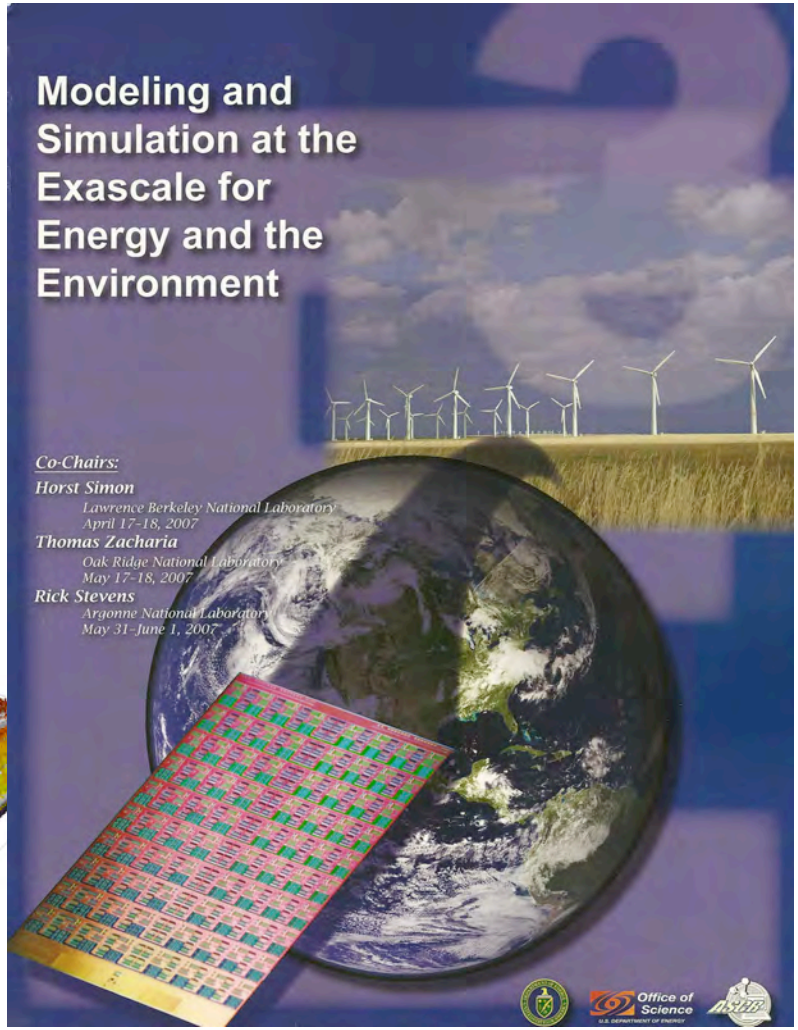
Argonne Leadership Computing Facility

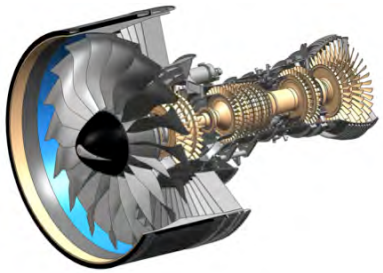
- Intrepid is fastest open science machine, third fastest overall
  - Peak: 557 Teraflops
  - Linpack: 450.3
- Configuration
  - 163840 cores
  - 80 Terabytes of memory
  - 8 Petabytes of disk storage
  - 10,000 volume tape archive

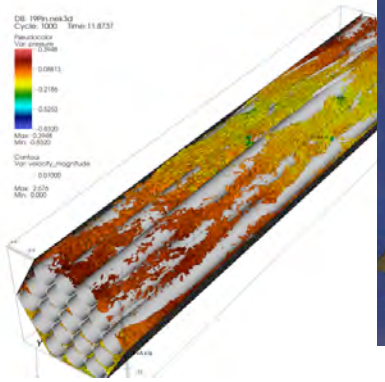
Exploring Exascale





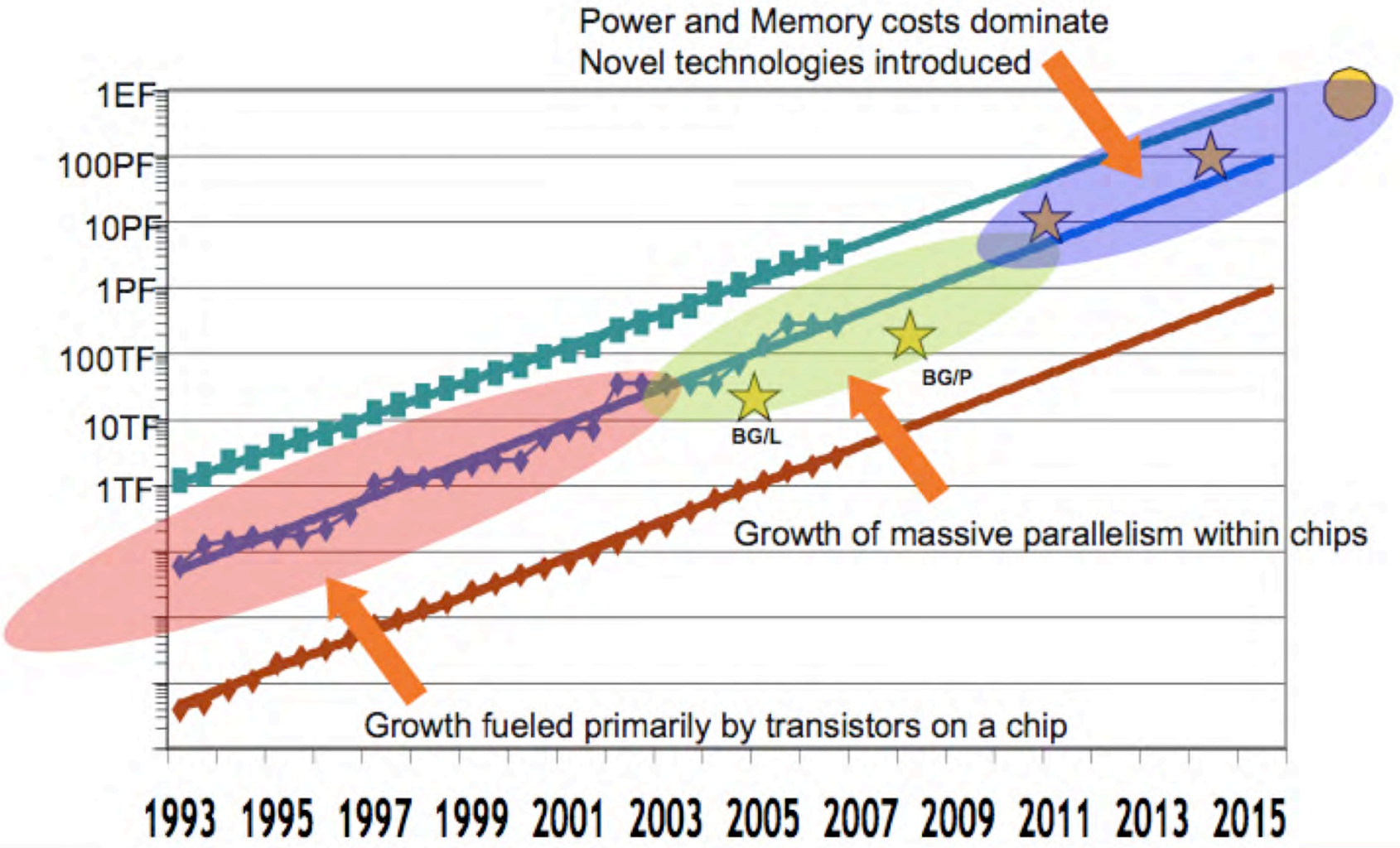






www.er.doe.gov/ASCR/ProgramDocuments/TownHall.pdf

Looking to Exascale



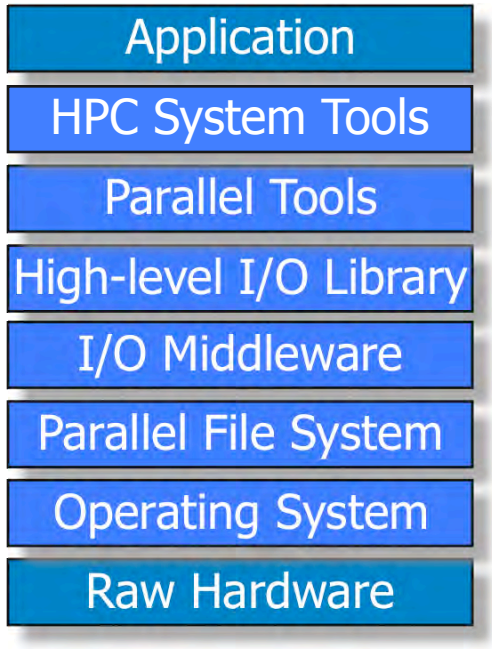
A Three Step Path to Exascale

Begin Full System Delivery (Yr)	2004	2008	2012	2015	2019
Design Parameters	**BG/L**	**BG/P**	**ONE**	**TWO**	**THREE**
Cores / Node	2	4	8-24	32-64	96-128
Clock Speed (GHz)	0.7	0.85	1.6-4.1	2.3-4.8	2.8-6.0
Flops / Clock / Core	4	4	8-32	8-32	16-64
Nodes / Rack	1024	1024	100-512	256-1024	256-1024
Racks / Full System Config	64	72	128-350	128-400	256-400
MB RAM/core	256	512	1024-4096	1024-4096	1024-4096
Total Power	2.5MW	4.8MW	8MW-20MW	20MW-50MW	40MW-80MW
Flops / Node (GF)	5.6	14	128-640	640-2000	2000-6000
Flops / Rack (TF)	5.7	14	200-400	400-1200	1600-4800
LB Concurrency	5.E+05	1.E+06	1M-2M	10M-100M	400M-100M
Full System					
Total Cores (Millions)	0.13	0.3	.3M-1.2M	1M-10M	4M-30M
Total RAM (TB)	33.6	151	2,000-4,400	3,000-10,000	5,000-25,000
Total Racks	64	72	128-350	128-400	256-400
Peak Flops System (PF)	**0.37**	**1**	**25**	**300**	**1200**

Linux on 100M Processors (Cores)?

- I agree:
  - The formulas always boil down to cost…
  - Don't buy a machine without threads
- Free (with some custom mods) is VERY attractive

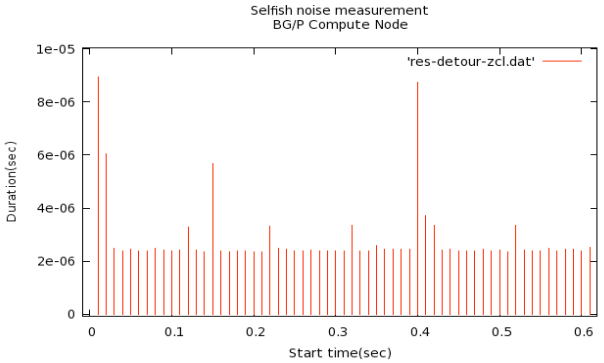
- Let's look at Linux
  - Its the de facto standard for SciComp.
    - *Number 1 platform target for HPC scientific tools*
  - Is extremely flexible across many scales
    - *From the cell phone to the top500*
  - Is High Productivity!
    - *UPC on Linux for BG/P… 5 minutes…*



The Question is not Why… but Why Not?

Possible "Why Not" Answers…

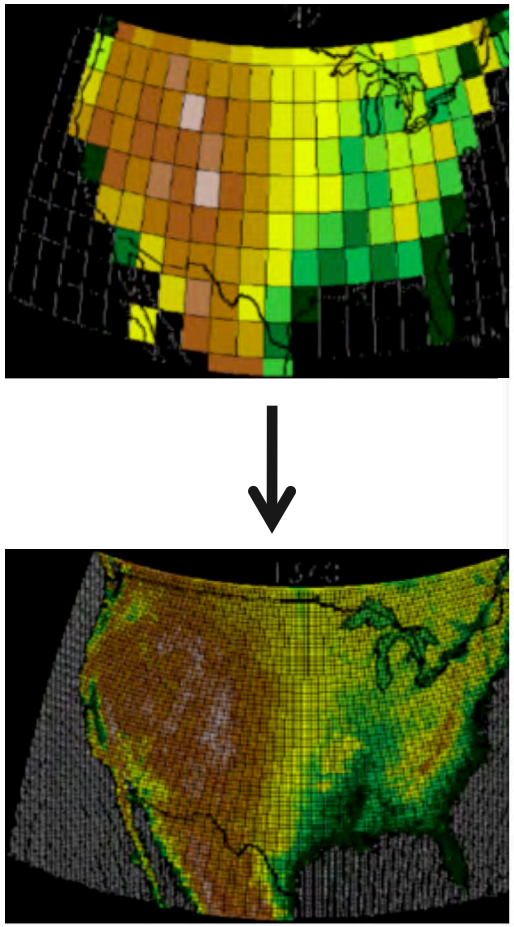
- Exascale Hardware could be wildly different from today's hardware. Linux will be too far away
  - Hmm, maybe. Good area for research!
  - Pico-joule energy management layer, TransMem, etc
- Linux is "Too Slow" or "Too Big".
  - Generally irrelevant -- HPC Codes use libraries heavily (libc, I/O, etc), but the kernel just handles simple low-level functions
- OS Noise!
  - Generally not an issue for well hacked Linux kernels
  - Will be irrelevant in a couple years anyway
- Memory management
  - Yes, this one is hard…



Why lock-step programming is dead

- Models are becoming more sophisticated

Old



New

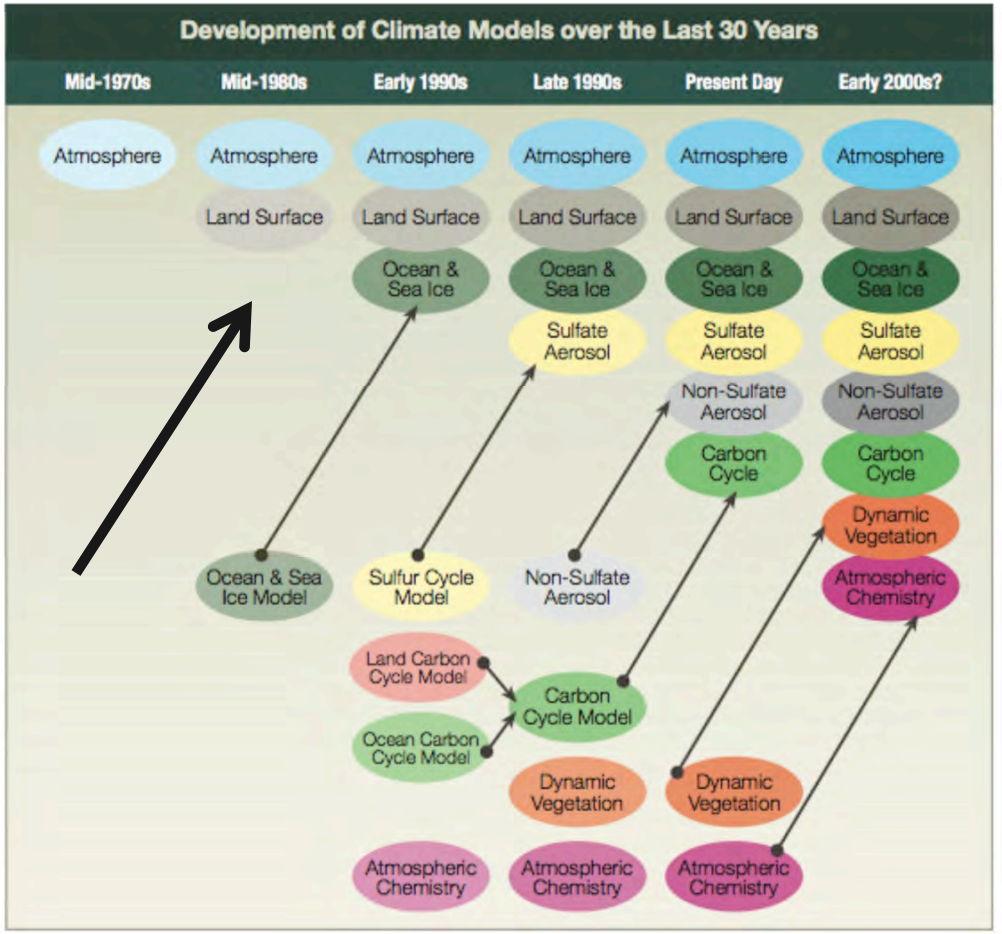
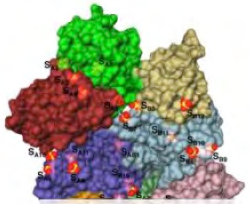


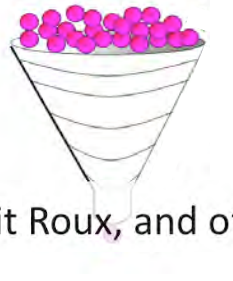
Figure 11: Development of Climate Models over the Last 30 Years. The development of climate models over the last 30 years showing how the different components are first developed separately and later coupled into comprehensive climate models. *Credit:* **CCSP Strategic Plan,** *Chapter 10 (2003).*

DOCK: Identifying Potential Drug Targets

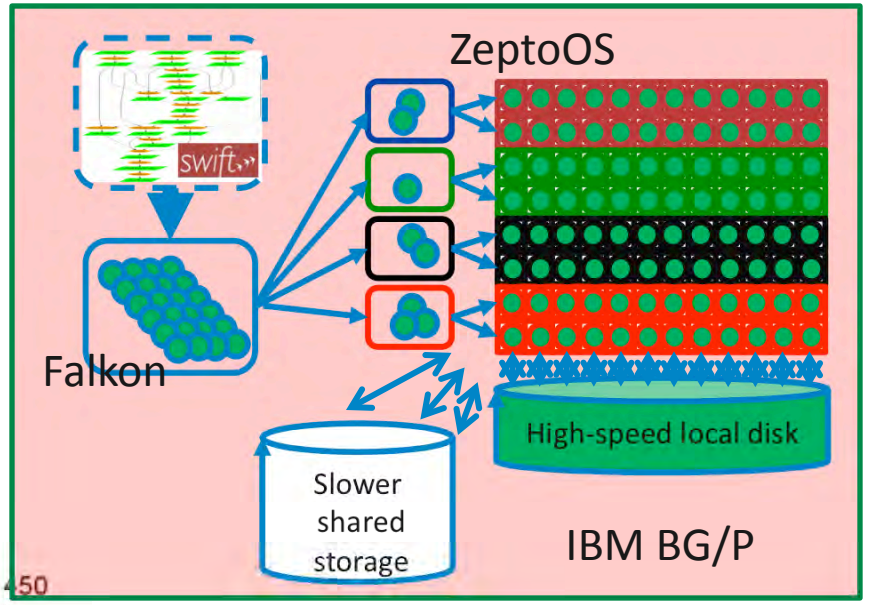
ZeptoOS + Falkon + Application

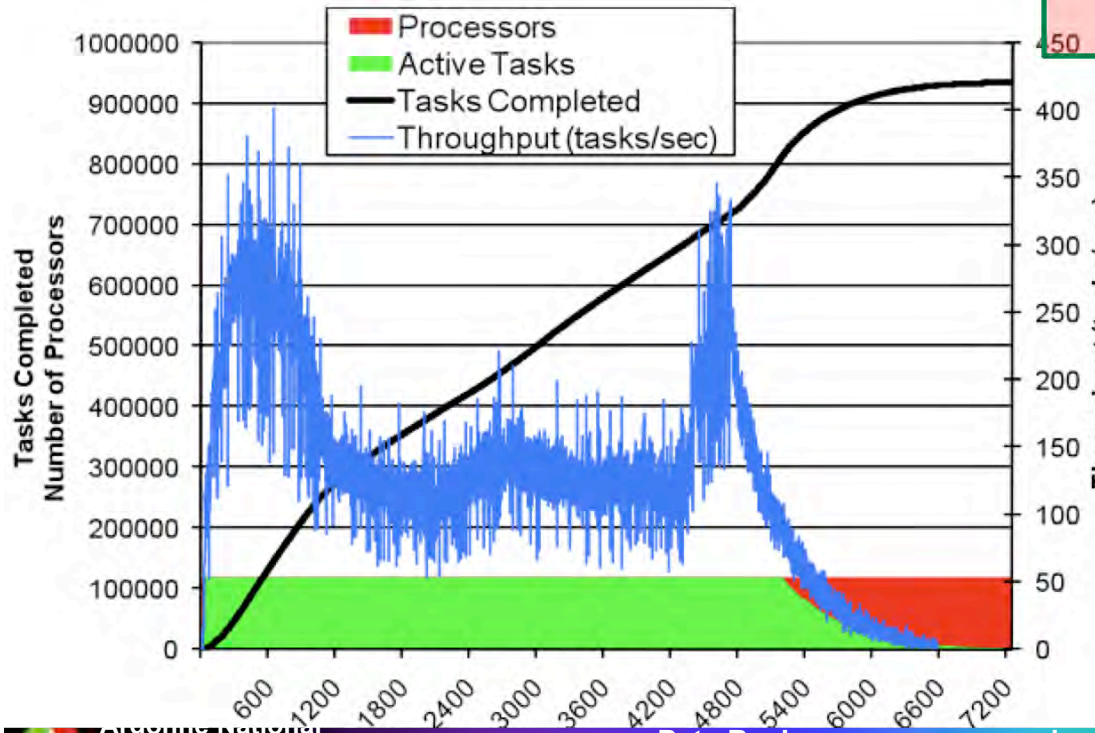
Protein target(s) x 2M+ ligands





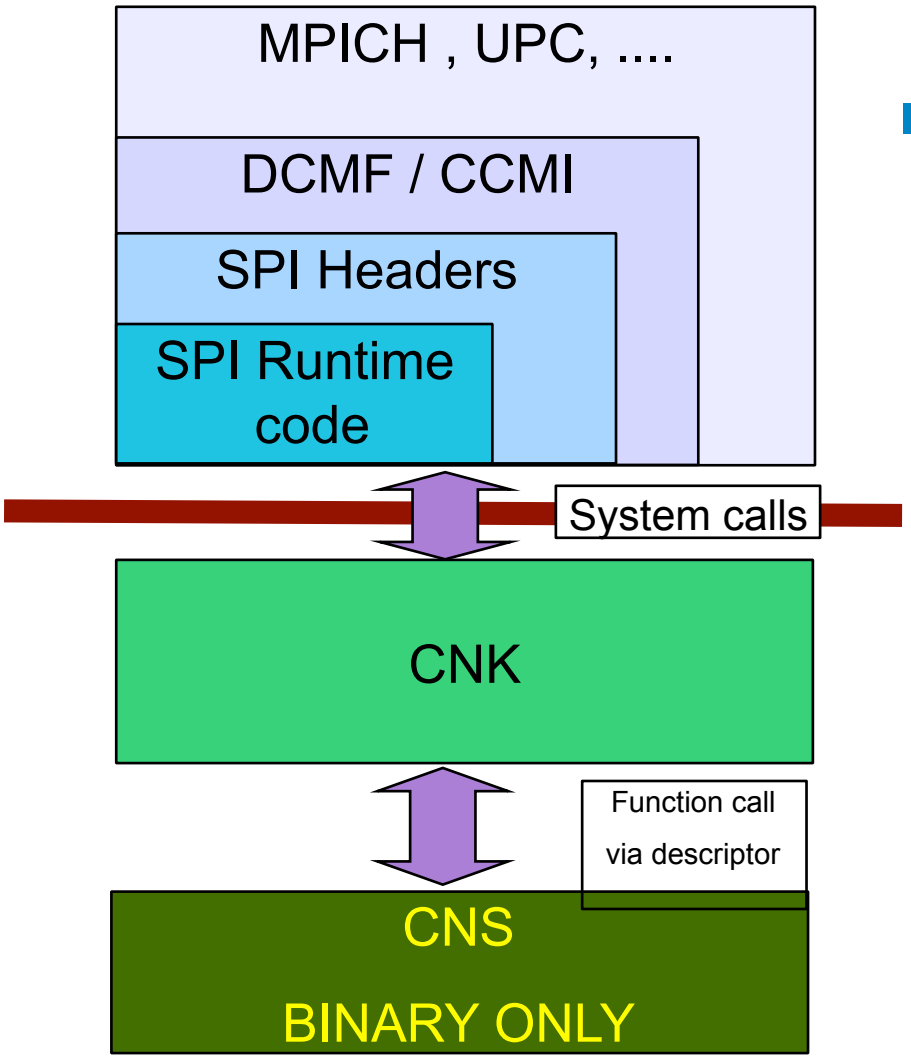
(Mike Kubal, Benoit Roux, and others)





- CPU cores: 118784
- Tasks: 934803
- Elapsed time: 2.01 hours
- Compute time: 21.43 CPU years
- Average task time: 667 sec
- Relative Efficiency: 99.7% (from 16 to 32 racks)
- Utilization:
  - Sustained: 99.6%
  - Overall: 78.3%

BG/P CNK Software Stack

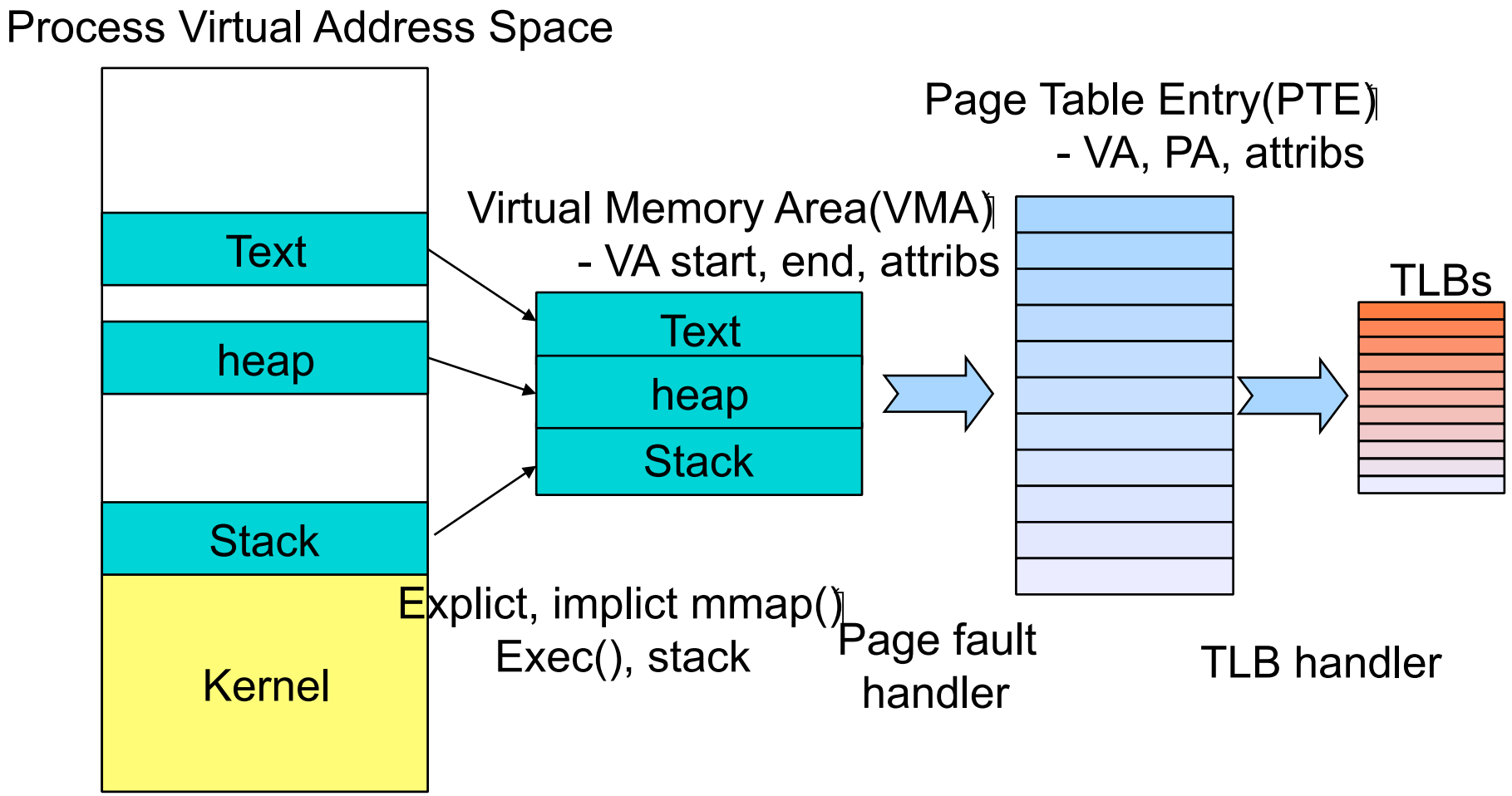


- CNS
  - Loaded at boot
    - *CNS is not library*
  - Functions call via CNS descriptor
    - *Torus DMA control*
    - *Interrupt control*
    - *Mailbox*
      - RAS, console msg
    - *etc*

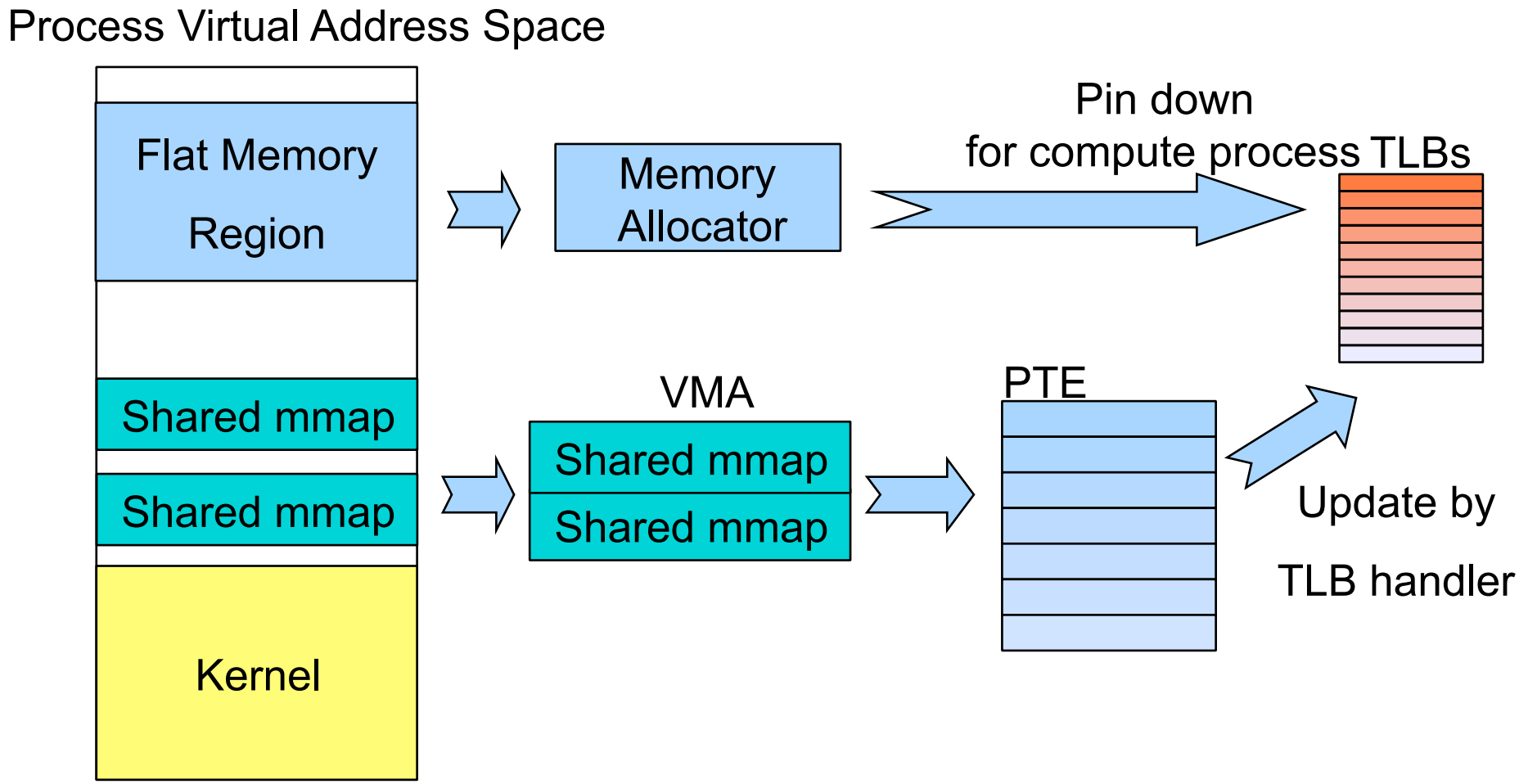
Compute Node Linux

- CN Linux Ramdisk
  - Similar to IO ramdisk
  - Embedded in kernel image
  - Basic Linux tools
  - Debug tools. i.e. strace, gdb
  - Zoid , Fuse client
- Remote log-in and file system
- Problems:
  - Linux Paging
    - *disadvantage on performance*
    - *Torus DMA requires physical contiguous memory*
  - BG/P software stack
    - *not designed for Linux !*
    - *i.e. CNK specific system calls*

Linux Standard MM



Hybrid approach
Flat memory and paging



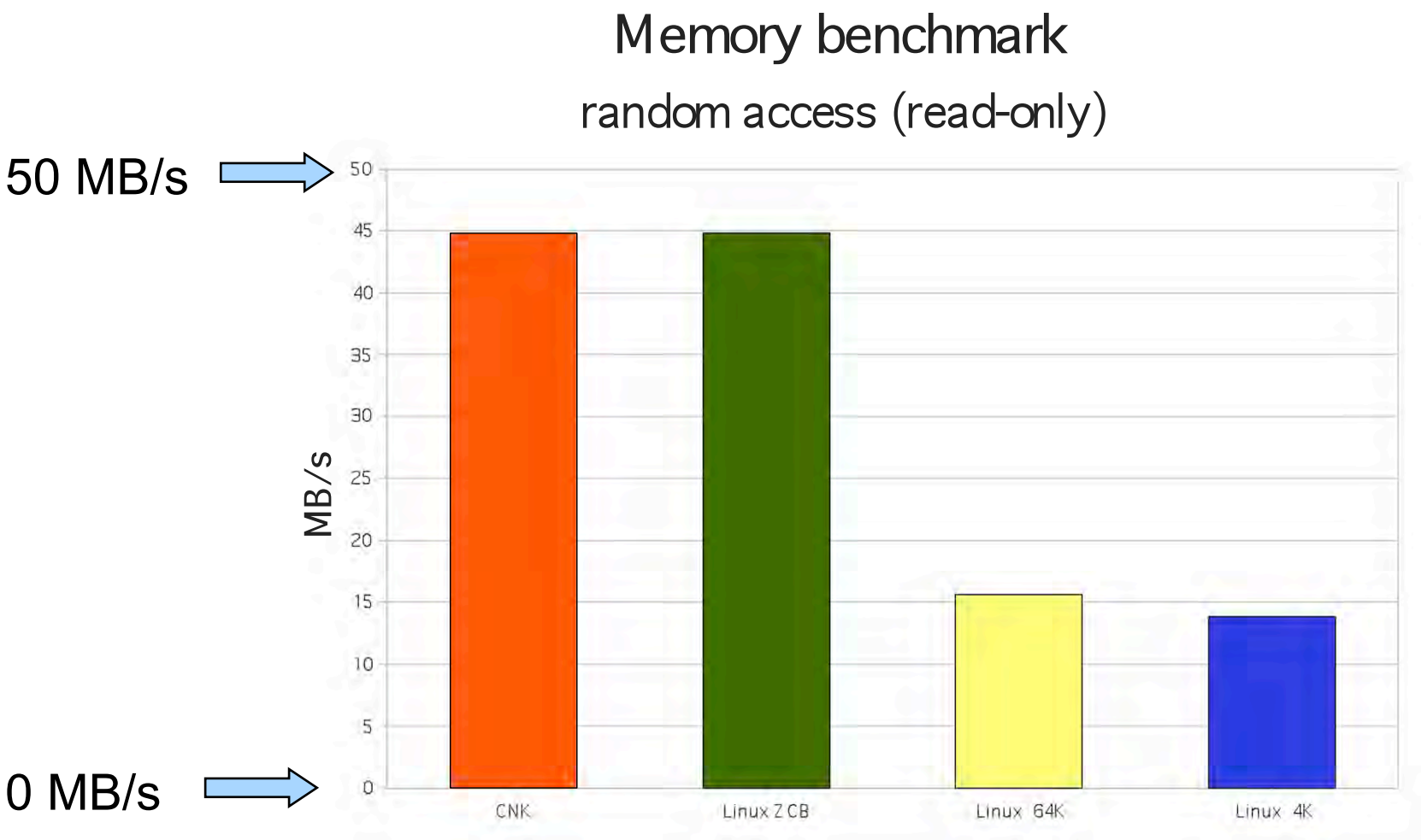
Flat memory mapping

- Flat memory region is reserved at boot time
  - Only for Zepto Compute Node binary (ZCB)
- It's transparent!
  - *No code modification, no re-compilation*
    - OS automatically loads flat mapping for compute process
  - *Enabled or disabled by the CUI tool*
    - Alternate flag in ELF header

```
# zelftool  ./memperf
zcb is off
# zelftool -e ./memperf

# zelftool  ./memperf
zcb is on
```

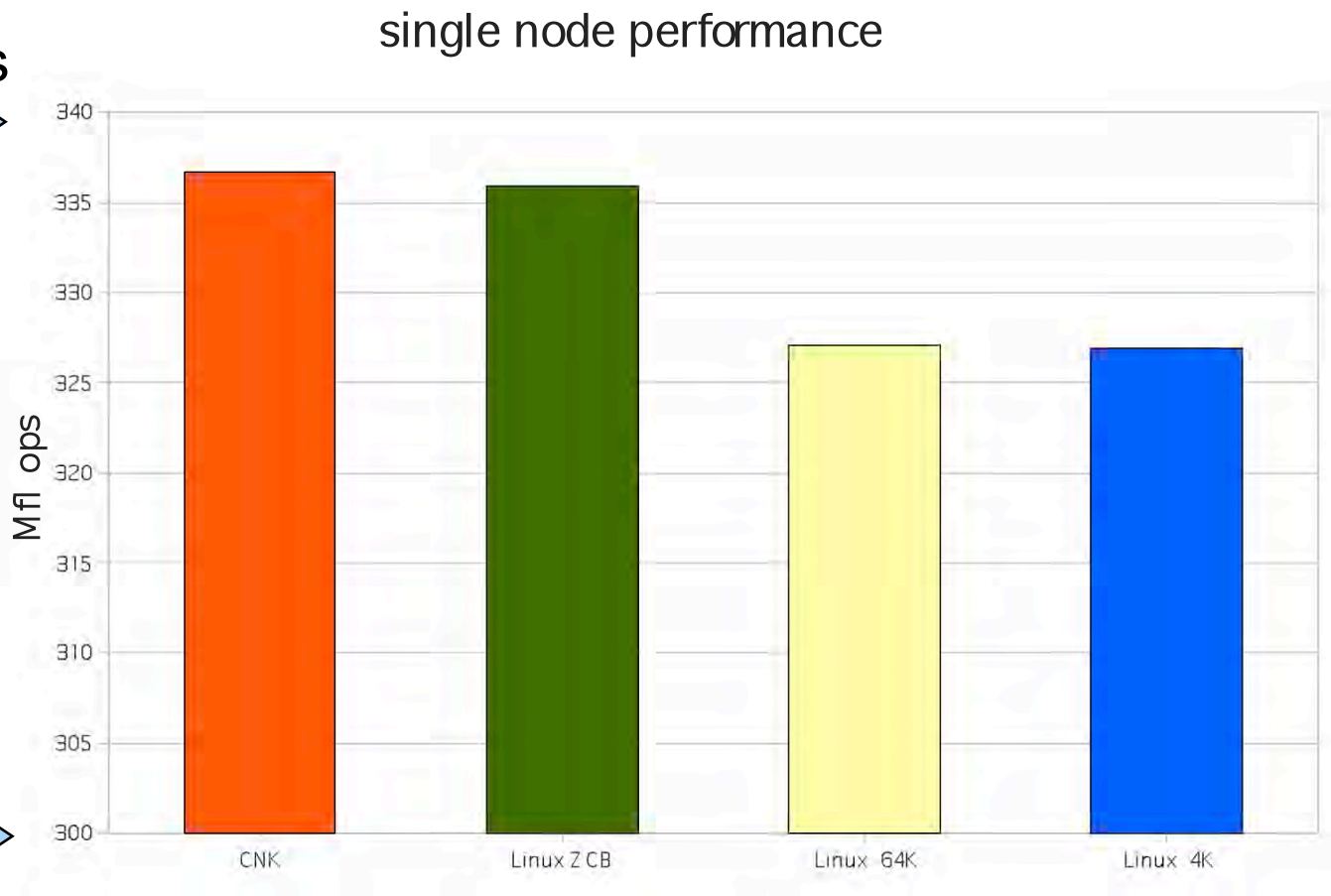
TLB stress benchmark results



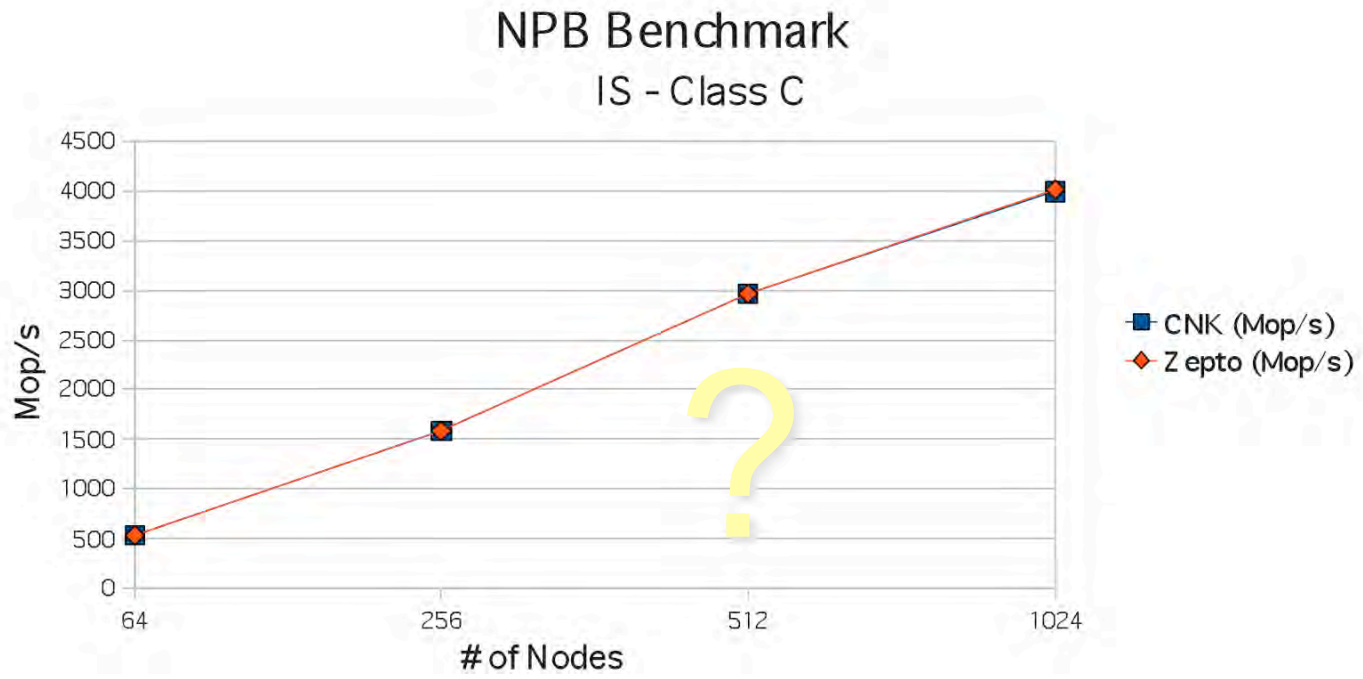
FFTW benchmark

340 Mflops

300 Mflops

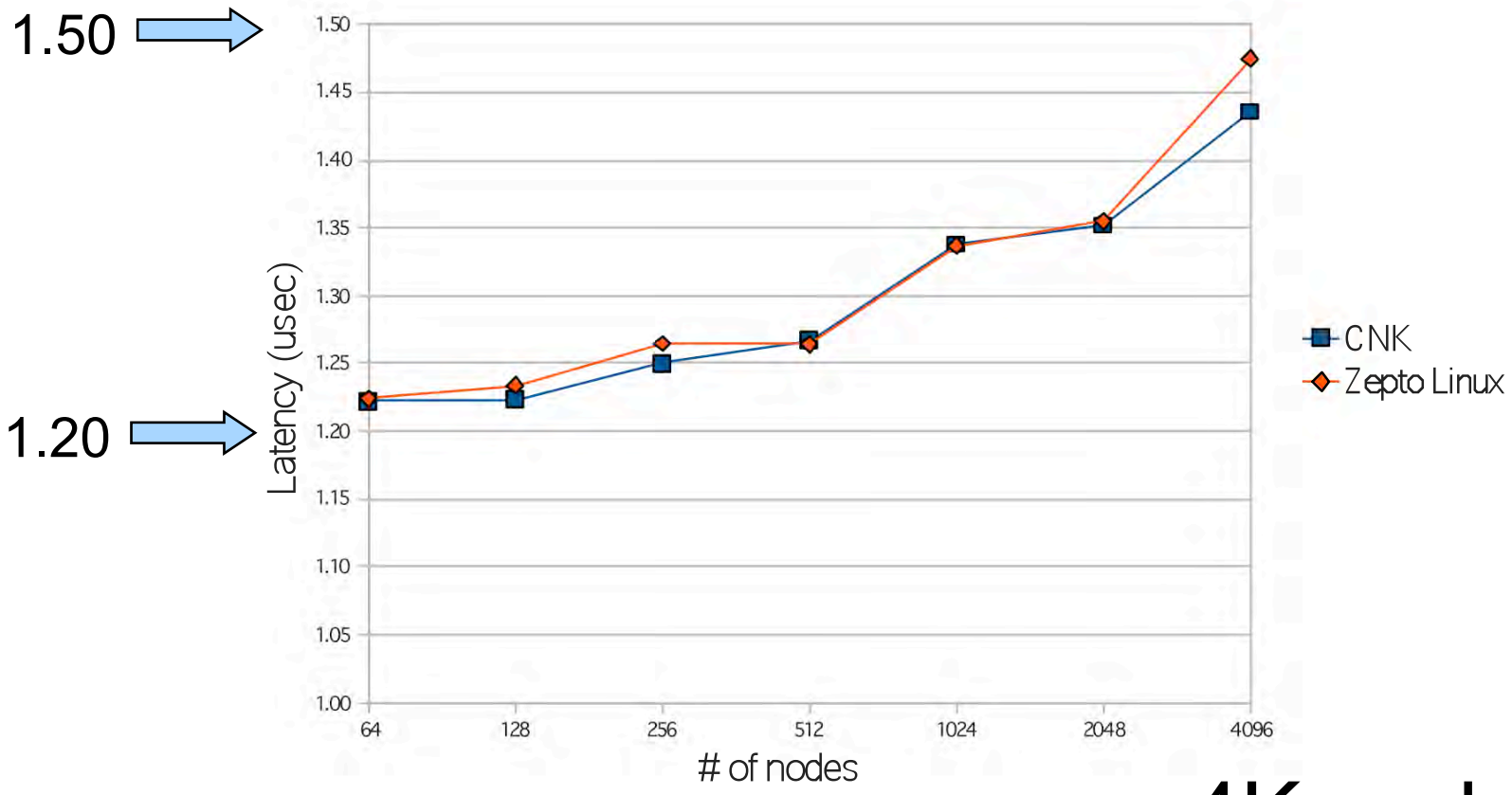


NAS Parallel Benchmark Result



- Performance difference < 1.0 %
  - Worst case: LU 0.7% slower
  - Best cast : IS 0.5% faster

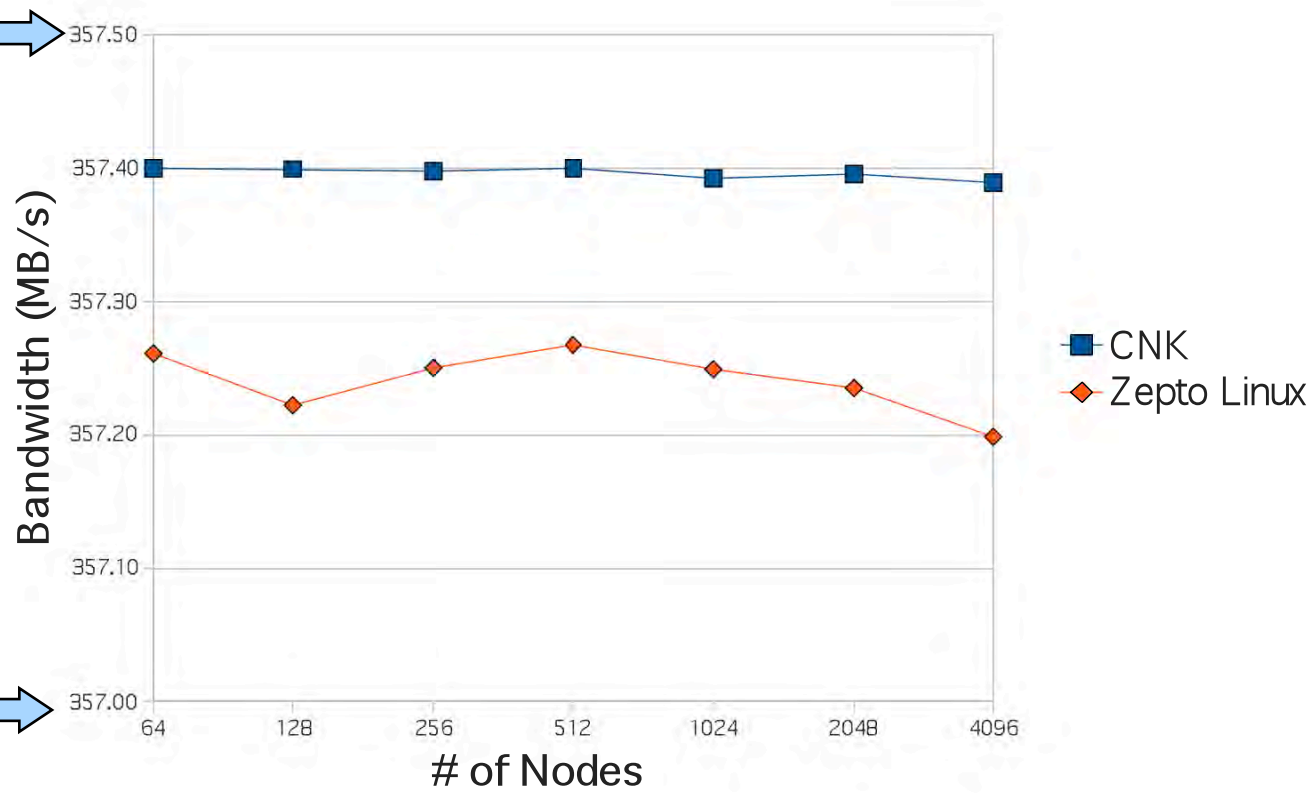
MPI Barrier Latency



4K nodes

MPI Send/Recv Bandwidth

357.5 MB/s

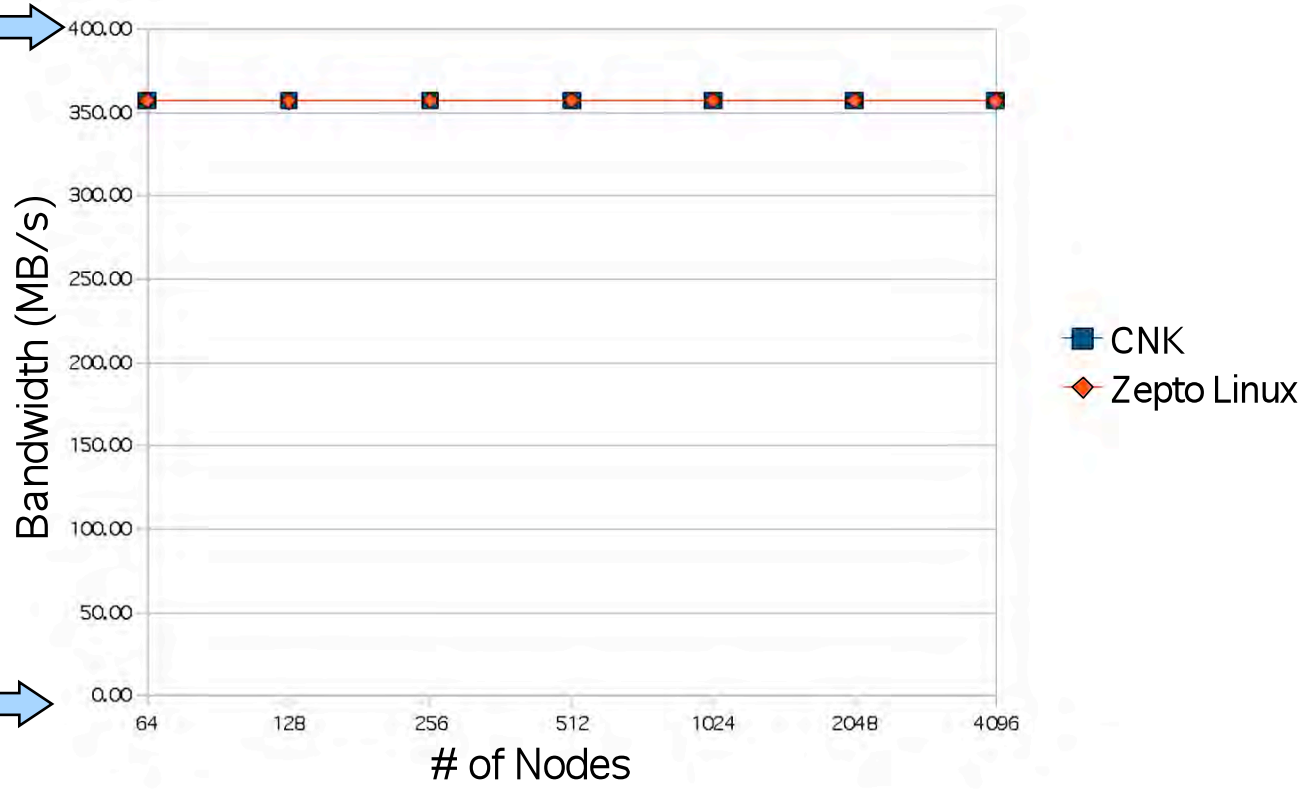


357.0 MB/S

Difference is about 0.05 %

MPI Send/Recv Bandwidth

400 MB/s



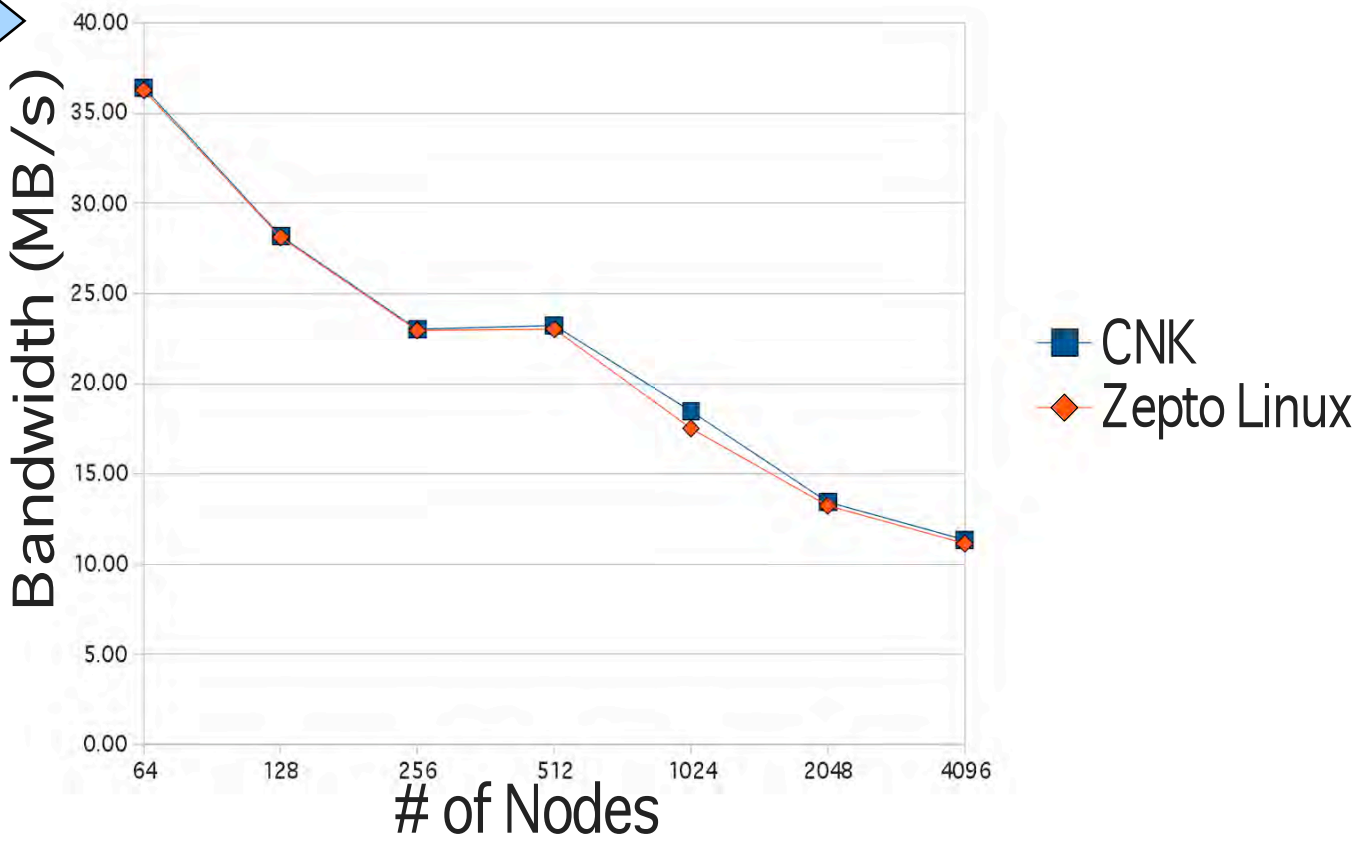
0 MB/S

Difference is about 0.05 %

MPI Allreduce MPI_OP Bandwidth

40 MB/s

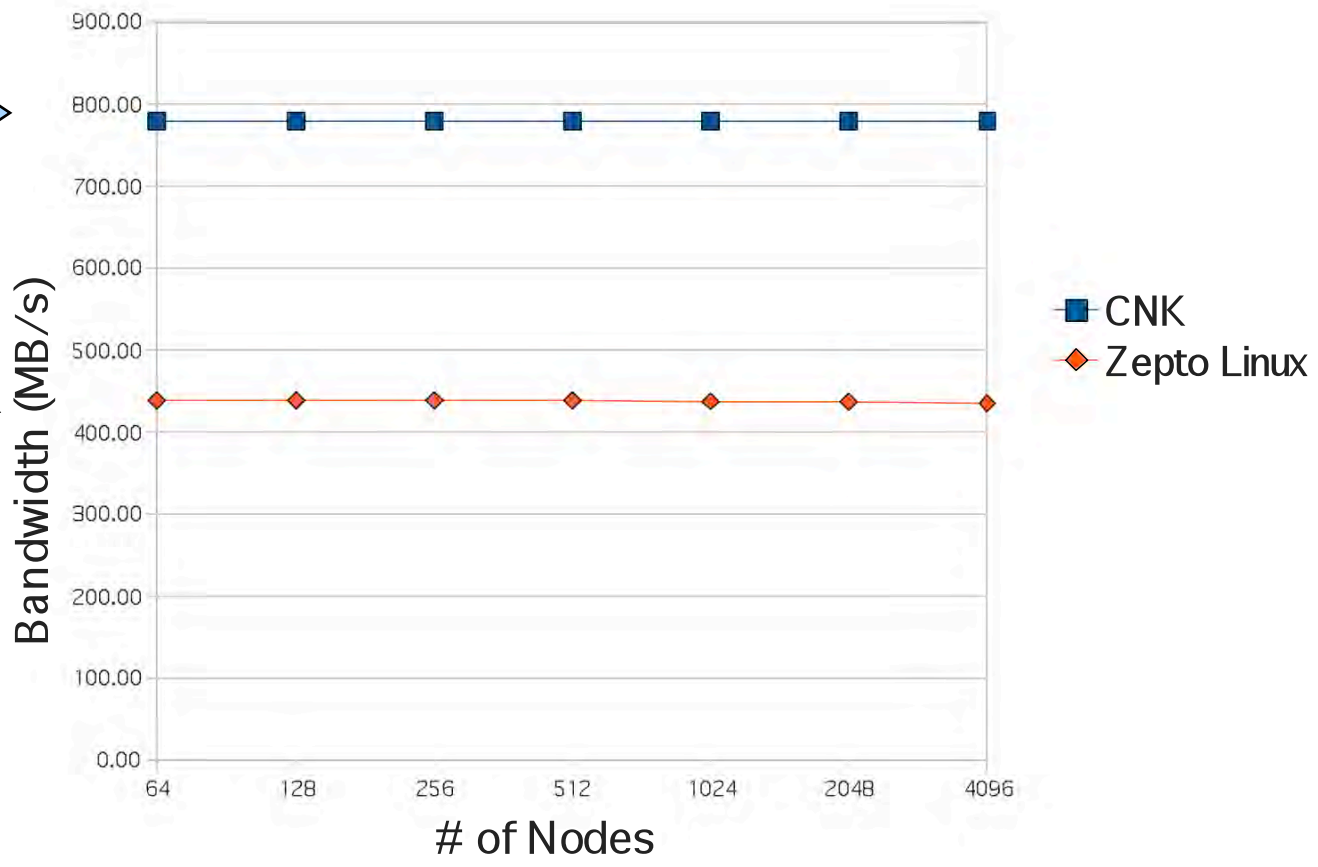
10 MB/s



MPI Allreduce MPI_INT Bandwidth

780 MB/s

440 MB/s



Revisiting "Why Not" Answers…

- Exascale Hardware could be wildly different from today's hardware. Linux will be too far away
  - Hmm, maybe. Good area for research!
  - Pico-joule energy management layer, TransMem, etc
- Linux is "Too Slow" or "Too Big".
  - Generally irrelevant -- HPC Codes use libraries heavily (libc, I/O, etc), but the kernel just handles simple low-level functions
- OS Noise!
  - Generally not an issue for well hacked Linux kernels
  - Will be irrelevant in a couple years anyway
- Memory management
  - Yes, this one is hard…