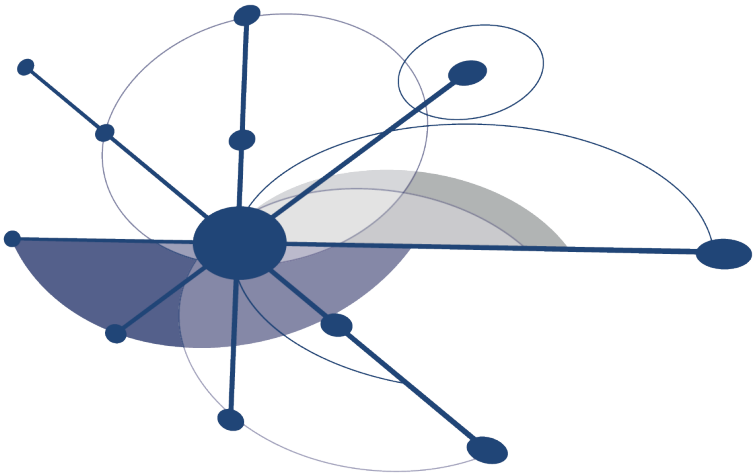


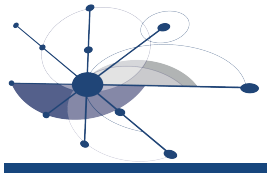
Center for Scalable Application Development Software: System Software

Peter Beckman (ANL)

SciDAC
Scientific Discovery through Advanced Computing

CScADS ZeptoOS Research

- Exploring performance improvements for system software on leadership-class multicore platforms
- Focus
  - memory management
  - I/O forwarding and job control
  - communication software stack
- Benefits
  - foster software research on leadership computing platforms
  - extend the usage of leadership computing platforms

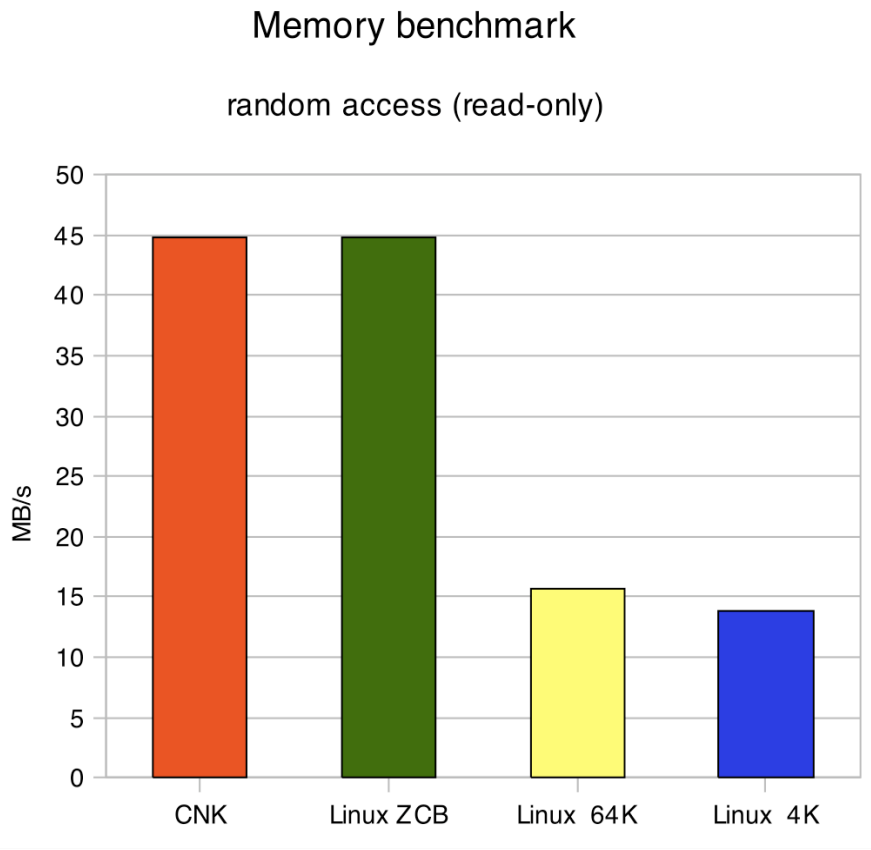


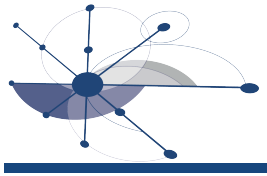
Motivation

- Memory Management
  - overhead of general-purpose paged memory (*not* OS jitter) key issue when running Linux on IBM Blue Gene
  - physically contiguous memory needed by some hardware devices
- I/O
  - 200K clients on current machines, millions on next generation—will file systems even be able to handle this?
  - Argonne's 557 TF Blue Gene/P (Intrepid):
    - 20% of the money spent on I/O
    - full memory dump takes over 30 minutes
  - I/O quickly becoming *the* bottleneck:
    - we need to make I/O as efficient as possible
    - flexibility

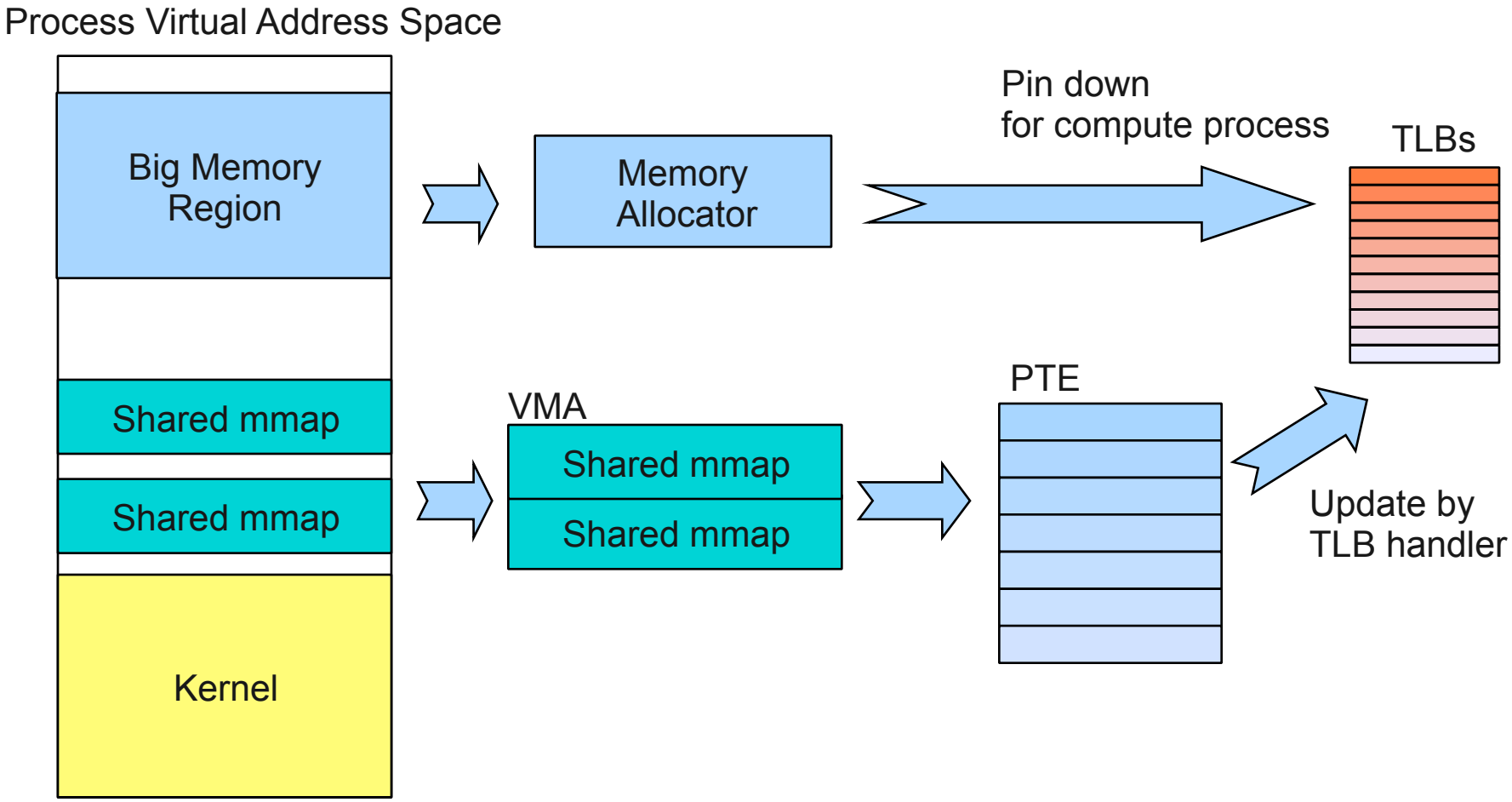
Memory Management on BG/P

- General purpose OS loses memory performance
  - worst case: standard Linux on ppc450 achieves only 20–25% of the theoretical memory bandwidth due to high cost of TLB misses
- Solution
  - introduced Big Memory management to Linux
  - enables a compute task to access memory without TLB misses



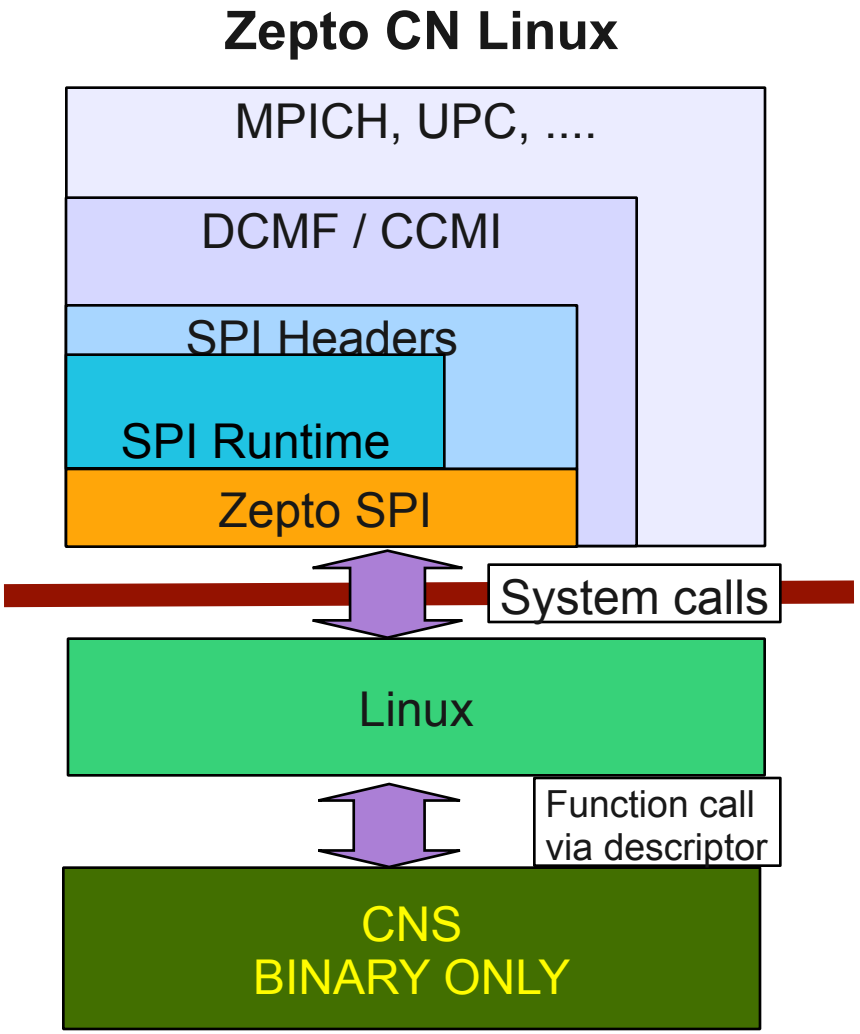


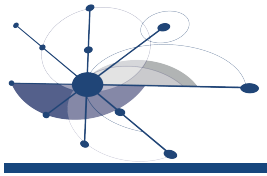
Our Approach – Big Memory



BG/P Compute Node Software Stack

- ZeptoOS Compute Node Linux
  - Big Memory for performance and torus DMA
- Deep Computing Messaging Framework (DCMF)
  - low-level communication layer that other communication APIs are built upon
- MPICH
- Unified Parallel C (UPC)



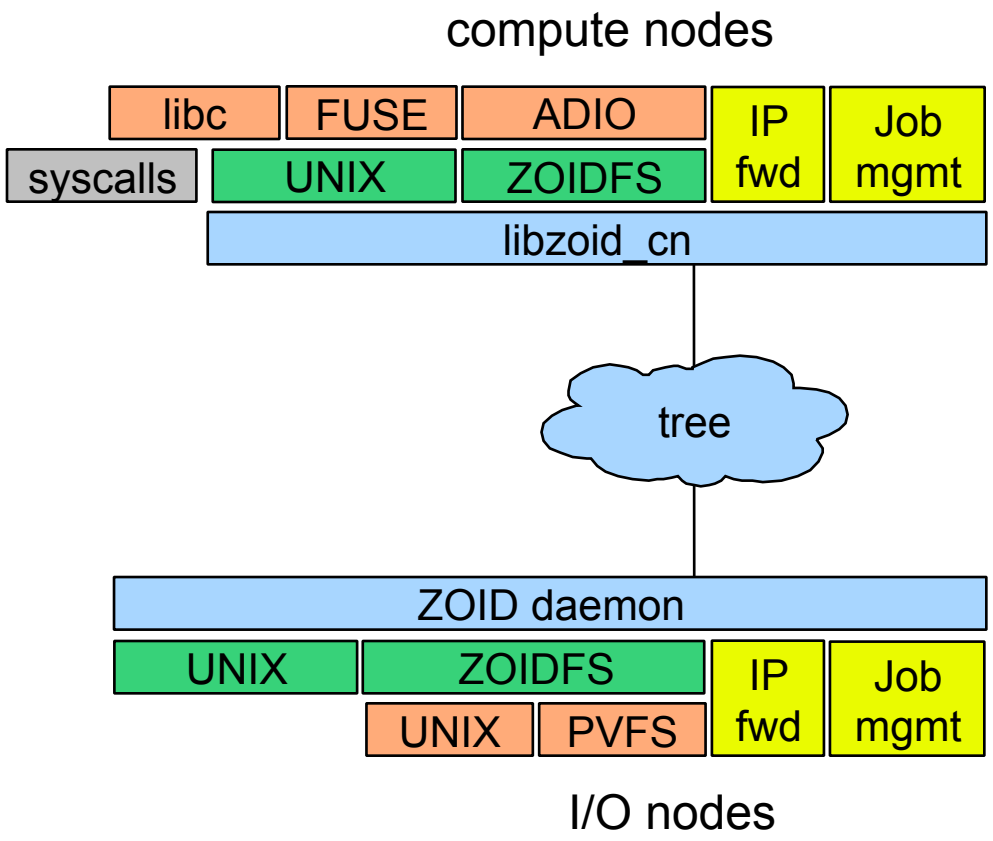


I/O Forwarding and Job Control

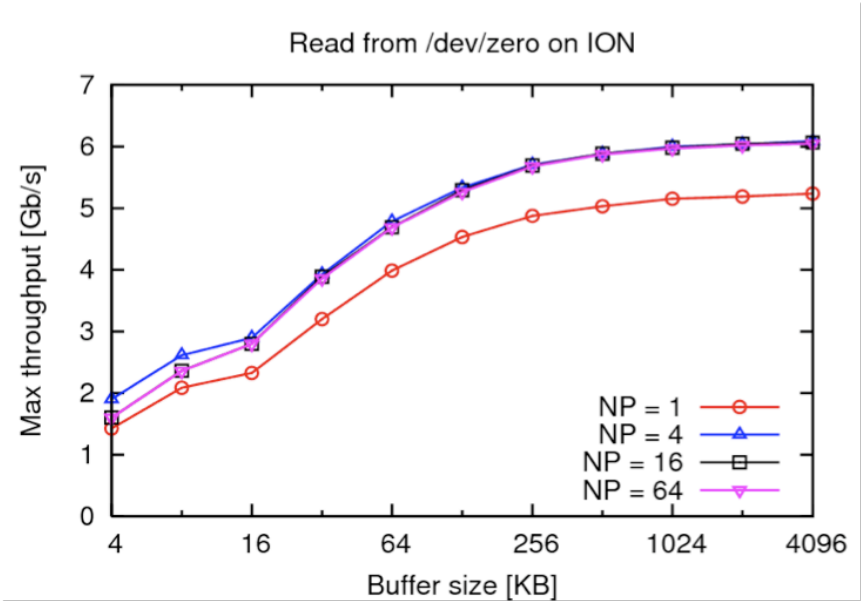
- ZOID (ZeptoOS I/O Daemon) provides
  - complete job management
  - file I/O and IP forwarding for Zepto Compute Node Linux
- Extensible through plugins
  - custom I/O forwarding APIs
    - e.g. file system client, communication layer
- Open, full source code available
  - enables independent computer science research
- Optimized performance
  - multithreading to hide latency
  - reduced context switching

ZeptoOS I/O Daemon (ZOID)

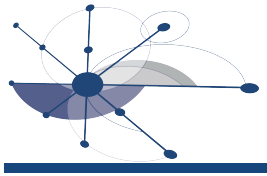
Architecture



Performance



(raw link bandwidth is 6.8 Gb/s)



ZeptoOS Results

Blue Gene/P Compute Node OS and I/O layer operational

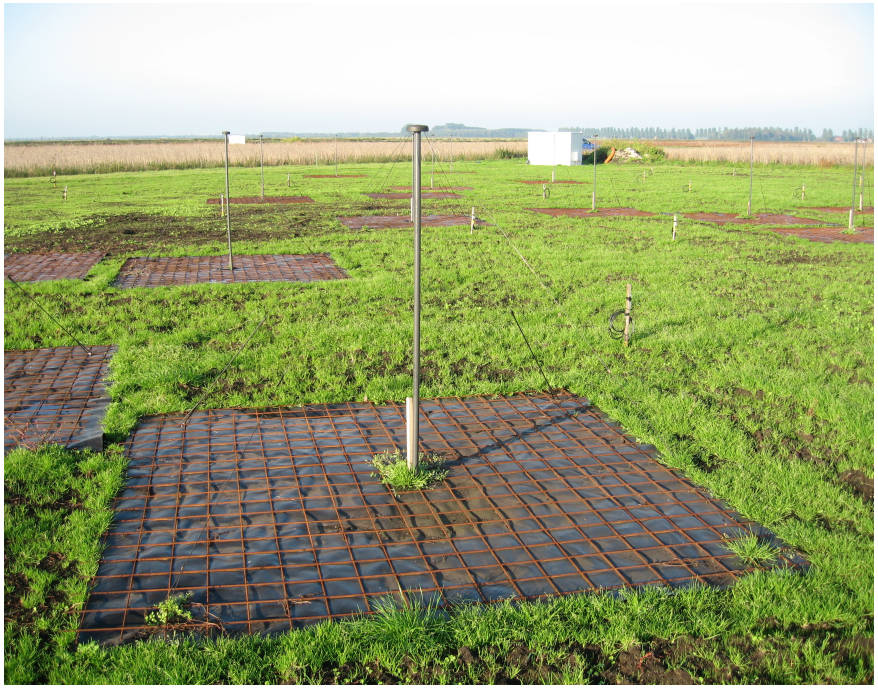
First prerelease for BG/P available, full release imminent

- Supports High Performance Computing (HPC) on BG/P
  - BG/P compute node software stack has been ported
  - MPICH is ready to use in SMP mode
  - Negligible performance penalty on NAS benchmarks
- Supports High Throughput Computing (HTC) on BG/P
  - Falkon task execution framework has been ported

LOFAR

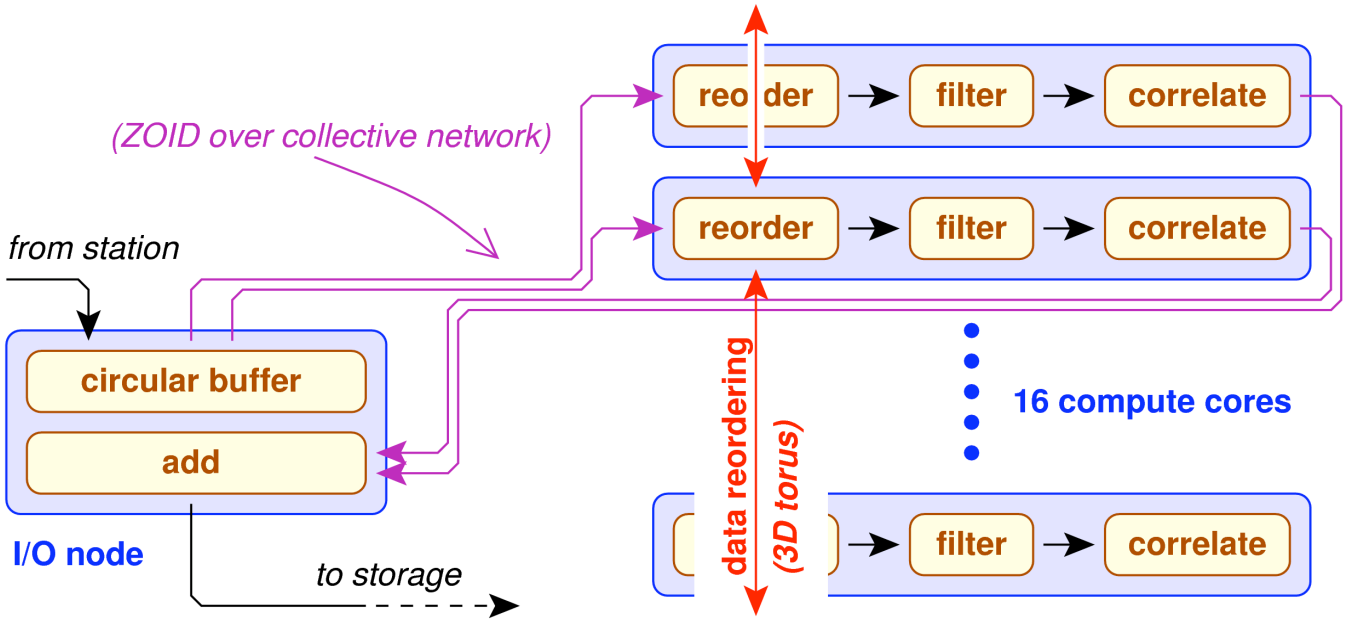
LOw Frequency Array

- Revolutionary radio telescope
  - no dishes
  - *O*(10000) receivers
  - omni-directional
- Central processing
  - real time
  - *Soft*ware
  - BG/L supercomputer



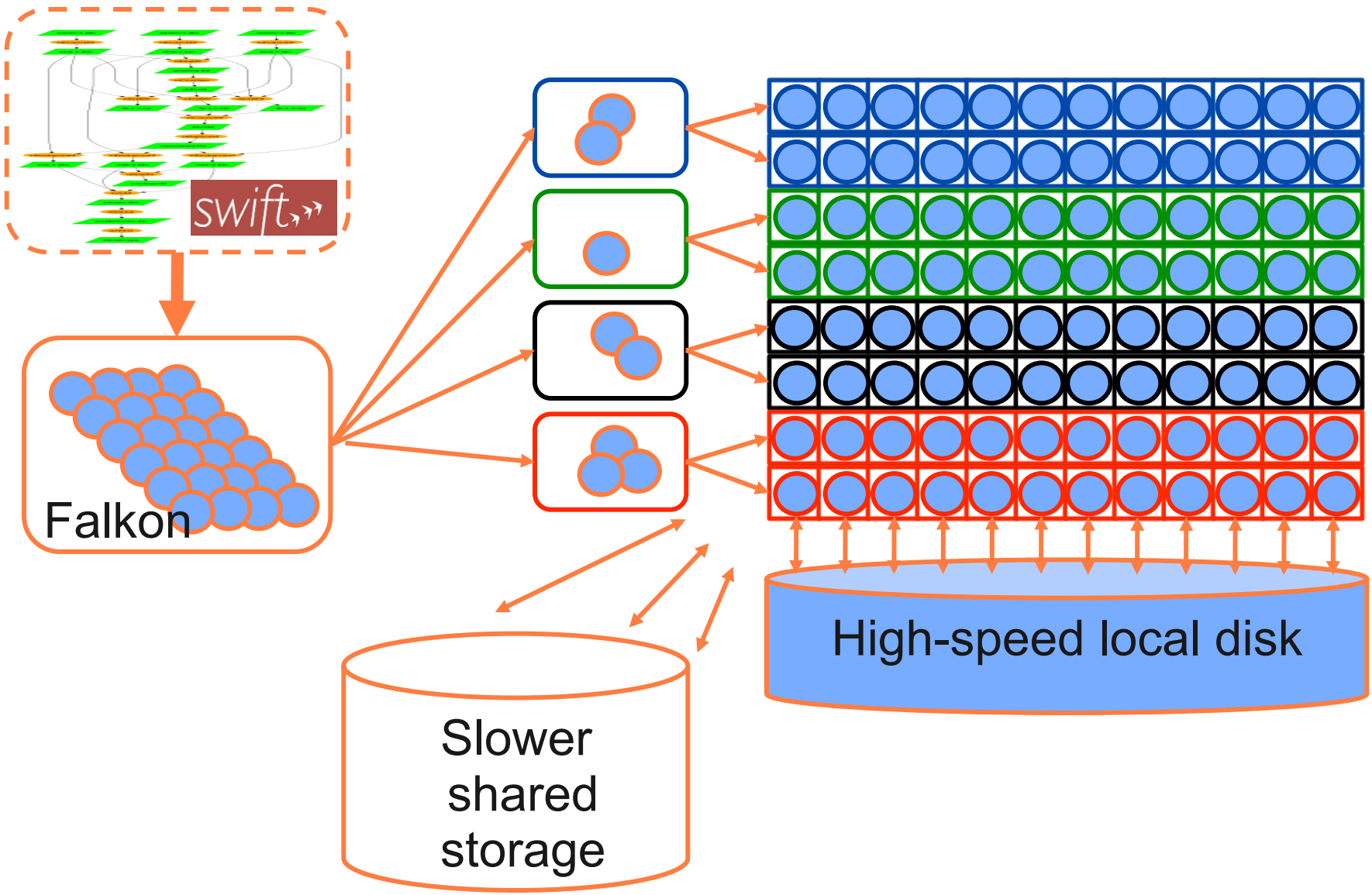
LOFAR Processing with ZOID

- Reorder, filter, correlate data
- Use zoid plug-in on I/O node

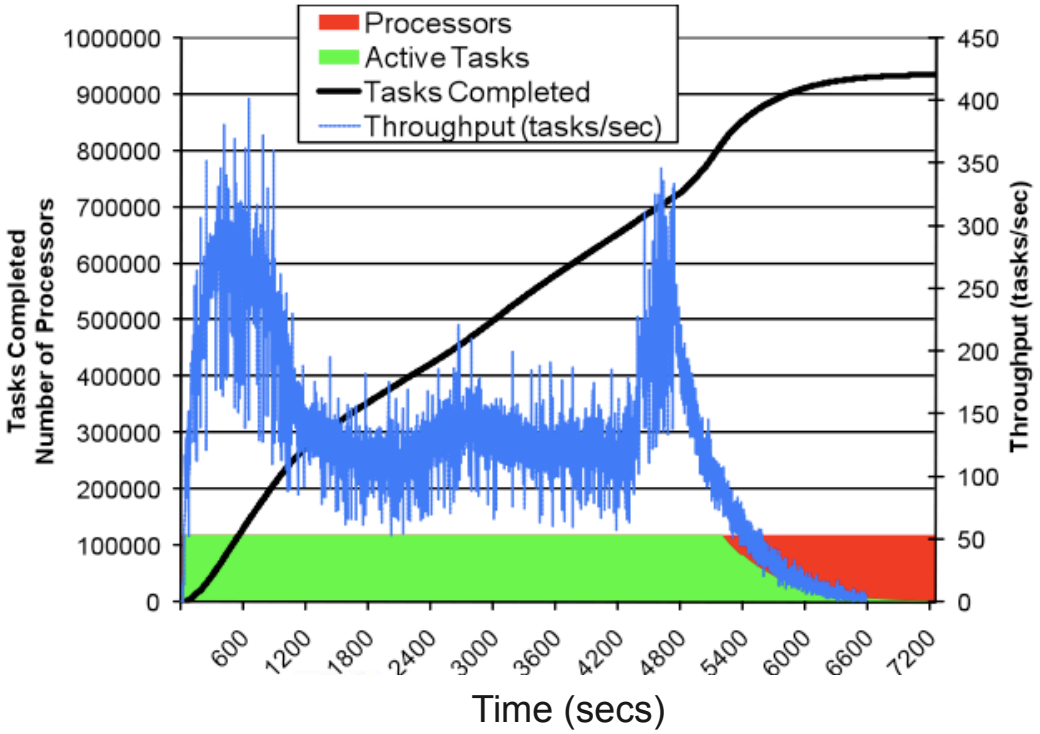


- Application on I/O node: no need for input cluster

Falkon: Managing 160,000 CPUs



DOCK: ~1M Tasks on 118,000 CPUs



- CPU cores: 118784
- Tasks: 934803
- Elapsed time: 7257 sec
- 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%

- GPFS
  - 1 script (~5KB)
  - 2 file read (~10KB)
  - 1 file write (~10KB)
- RAM (cached from GPFS on first task per node)
  - 1 binary (~7MB)
  - static input data (~45MB)