Software Updates and Ideas for CN5000

Douglas Fuller, VP Software Engineering
ExaComm at ISC 2024

CORNELIS™ NETWORKS

About Cornelis Networks

Provider of open, intelligent, high-performance networking solutions designed to accelerate the world's most demanding AI and HPC applications

Omni-Path CN5000

2024

Components

400G Foundational Adapter (PCIe Gen 5)
400G 48p Edge Switch
400G 576p Director Switch (air & liquid)

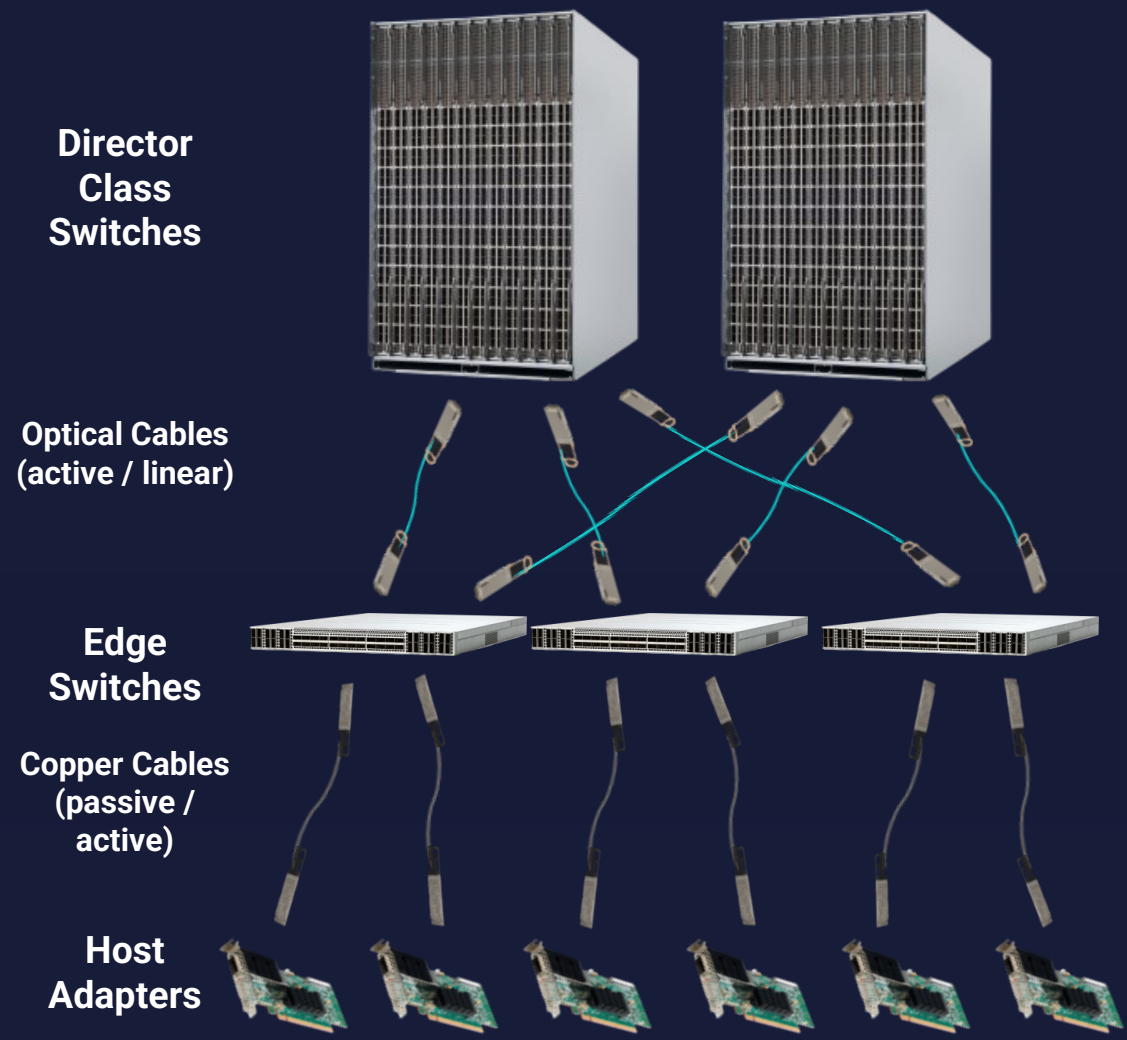
400G DACs up to 3m
400G ACCs up to 5m
400G LPOs up to 30m
400G AOCs up to 100m

Scalability

13.8K node full bisection bandwidth tree
27.6K node half bisection bandwidth tree
90K node Dragonfly
330K node Megafly

Additional Features

Telemetry-based dynamic adaptive routing
Advanced congestion control per service class
Virtualization support (SR-IOV)
Air & liquid cooling



Illustrative Data Center Diagram

Omni-Path Software

- Switch firmware and software
- Fabric Manager (opa-fm)
  - Subnet manager
  - Performance manager
  - Fabric executive
- Linux kernel driver (hfi1)
  - In-tree
  - Supports OPA100 and CN5000 HFIs
- Omni-Path Express (OPX)
  - libfabric provider
- User tools (Fast Fabric)
- SST / libfabric integration

Kernel API Changes

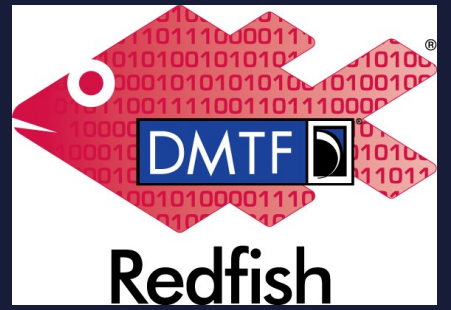
- SR-IOV implementation
- Driver remains the same (hfi1; supports both generations of devices)
- Some features require simple API changes (multiple ports were contemplated earlier)
- Maintain uverbs for now, may change this in the future
- Driver communication needs a file descriptor (DMA engines, etc.)
  - Might switch to uverbs for this
- `io_uring()`
  - We've been experimenting
  - Slow adoption from distros
  - Will move when we're confident
- OPX libfabric provider already has these changes
- LKML RFCs imminent!

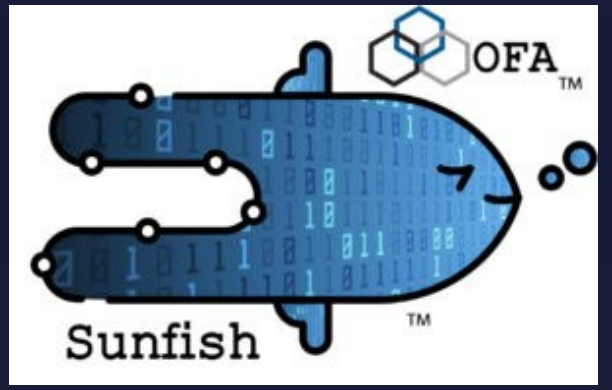
libfabric

- A software project of the OpenFabrics Interfaces Working Group (ofiwg)
- Communication API for high-performance networking
- Maps application semantics to underlying hardware features
- Providers define fabric services in a structured way
  - One/two-sided messages, counters, barriers, built-in collectives, etc.
- API users (like MVAPICH!) then consume these services with zero overhead
- Open source, community developed, widely used
- Our newest provider (`prov/opx`) will support both devices

Network Health and Monitoring

- Performance Manager (PM) and Performance Management Agent (PMA)
  - FM components that continuously monitor fabric health and performance
  - Tools for users to perform direct queries (API coming soon)
  - Avoids sweeping all devices manually for performance/status updates
- New switch health and management API
  - Based on DTMF Redfish and OFI Sunfish
  - Already some software that plugs into this
- Congestion telemetry
- Lots of data available programmatically
  - We'll still provide the usual tools
  - Collaborating with the community to cross-pollinate more ideas
- OpenBMC-based switch firmware allows for customization





New features for messaging libraries (or applications!)

User-controlled routing	Header field determines minimal, dispersive, or fine-grained adaptive routing
Receive-side matching rules	Configurable, hardware-implemented packet inspection Delivers matching packets to a particular receive context
Receive cut-through	Land packets with zero buffering
Receive bypass	Allocate large portions of contiguous host memory to land continuous flows

CORNELIS™
NETWORKS

Let’s Talk!

dfuller@cornelisnetworks.com