MUG 2024

CN5000 HW Deep Dive

Dennis Dalessandro - Cornelis Networks

CORNELIS™ NETWORKS

Notices and Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH CORNELIS NETWORKS PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN CORNELIS NETWORKS'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, CORNELIS NETWORKS ASSUMES NO LIABILITY WHATSOEVER, AND CORNELIS NETWORKS DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF CORNELIS NETWORKS PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. CORNELIS NETWORKS PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.

Cornelis Networks may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Cornelis Networks reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice. Roadmap not reflective of exact launch granularity and timing. The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Any code names featured are used internally within Cornelis Networks to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Cornelis Networks to use code names in advertising, promotion or marketing of any product or services and any such use of Cornelis Networks' internal code names is at the sole risk of the user.

All products, computer systems, dates and figures specified are preliminary based on current expectations and are subject to change without notice. Material in this presentation is intended as product positioning and not approved end user messaging.

Performance tests are measured using specific computer systems, components, software, operations, and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Cornelis Networks technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration.

Cornelis, Cornelis Networks, Omni-Path, Omni-Path Express, and the Cornelis Networks logo belong to Cornelis Networks, Inc. Other names and brands may be claimed as the property of others.

CORNELIS™ NETWORKS

Part 1: Introductory Material

Do we still need an intro?

- Spun out of Intel's OmniPath division a few years ago
- Many large deployments of 100Gbps OPA
- Coming later this year OPA 400 aka CN-5000

CORNELIS™ NETWORKS

Who am I? Why are you listening to me?

- Manage the Kernel team
- Working on OPA technology for over a decade
- Working with RDMA for 20+ years
- 2004 Graduate of OSU
- My son starts starts his time at OSU this week
- Formerly a researcher in this very building
- I like coming to talk to you folks!

What is CN5000?

- CN5000 is next generation fabric solution
- Consists of adapters, switches, cables, and software
- We will focus on how the HW works from SW point of view
- Why?
  - That's what the host sees
  - SW is what people interact with
  - Most important to projects like MVAPICH
  - *(and it's what I know about!)*

How does this help you?

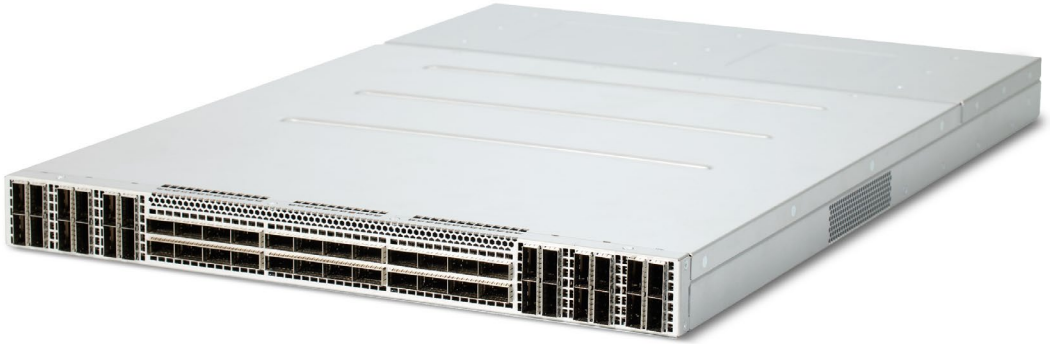
- Understanding HW means better SW
- Better SW means more breakthroughs
- CN is a fully committed upstream development organization
  - Which means we embrace community and change
- We are talking about HPC here one size fits-all does NOT APPLY

Now that I have your attention

- A brief product tour...

Switches

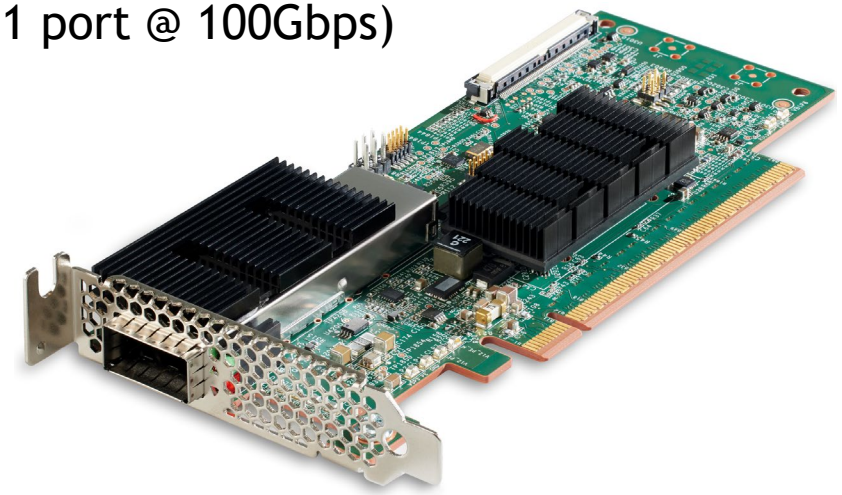
- Edge Switch
  - 48 Ports @ 400Gbps
  - Air, hybrid, liquid cooling available
- Director Switch
  - 576 Ports @ 400Gbps
  - Air and liquid cooling available
- https://www.cornelisnetworks.com/solutions/cornelis-cn5000/



CORNELIS™ NETWORKS

Adapters - Host Fabric Interface (HFI)

- The really cool stuff!
- PCIe Gen 5
- Low profile
  - Smaller is better. Have you tried to cram a GPU into a server?
- Air or indirect liquid using heat pipe from ASIC to server cold plate
- 1 or 2 Fabric ports @ 400Gbps (OPA-100 is 1 port @ 100Gbps)
- > 65 billion packets per second
- Latency as low as <1us



Software Overview

- Kernel
- Fabric Management
- Fabric Tools
- Messaging/Middleware

Kernel Support

- Long live hfi1
  - Despite the name we are continuing with the same driver
- Continues open source, upstream first development practices
  - Already in Kernel and in distro for OPA-100
  - Work closely with distros to ensure they have the best version of our software
- Upstream changes landing on the mailing list soon
  - Major changes (other than supporting a new chip which is hard enough)
    - Re-worked Management interaction with embedded CPORT
  - Supports multiple fabric ports
- Will support Nvidia and AMD GPU
  - Not upstream due to license issues
  - Still open source though, available on our GitHub

Fabric Management

- OPA FM package already in distro
- Fully open source and engaged fabric manager
- Robust routing algorithms
  - Congestion avoided via fine grained dynamic adaptive routing
  - State of the art congestion control
- In distro tooling

Messaging Support

- psm2 is riding off into the sunset
- Libfabric (OFI) is the non-proprietary way forward
  - OPX is our native provider
  - Designed around our HW capabilities
  - Highly optimized code
    - Boosts performance even on OPA-100
  - Active development
  - Open source so adaptable and extendable to your needs!

Hardware Details

- I focus on the adapter
- I can provide contact info for switch questions

Technical Highlights

- OPA-100 (100Gbps) adapter ASIC is known as Wolf River (WFR)
  - Name leaked from Intel long ago – old news
- CN5000 (400Gbps) adapter ASIC is known as Jackal River (JKR)
  - Flat out telling you what it is – because it's in the code and code names don’t matter
- Continue to take advantage of 16 DMA Engines
  - DMA Engines bring data into the card avoiding CPU copy
  - These are the large data transfers
- PIO (Programmed IO) capability increased
  - 160 contexts available in WFR
  - 240 contexts available in JKR
  - Memory increased 1MB to 4MB
- Full 16B packet type support in HW, as well as 9B
  - WFR only supported 9B in HW
  - 16B enables adaptive routing, larger LIDs (24bit vs 16)
- PKey table increased from 16 to 1024
  - Needed for MLS SELinux support

Highlights continued...

- Receive descriptors (how we land packets) have increased
  - From 65536 in WFR to 131072 in JKR
- Supports 8 VLs for data plus VL15 for mgmt
- PCIe SR-IOV support (lots of code changes coming for this!)
  - Dual loopback ports for SI to SI packet communication
- Integrated CPORT processor
  - Handles fabric mgmt (MAD packets in FW now, not in SW!)
- Receive Side Matching
  - Deeper packet inspection
  - More rules, from 4 to 32
  - Lots of ideas floating around in my head for these!
  - Your ideas welcome too! What could you do with them?
- Lots of other bells and whistles in the HW!

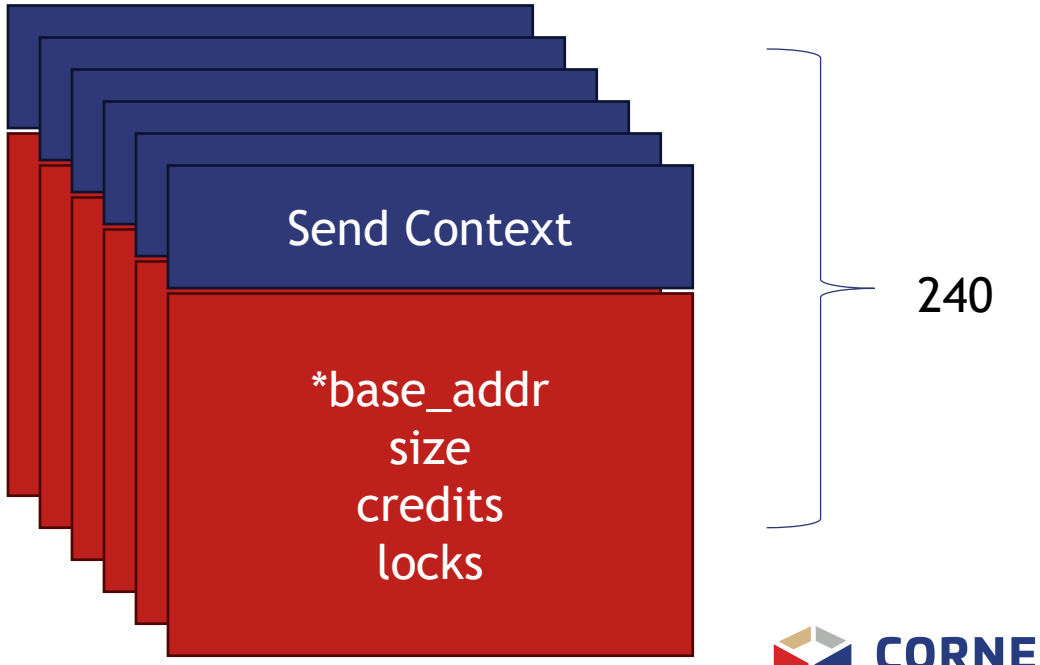
Part 2: HW and SW Details

Programmed IO aka PIO

- Uses CPU to copy data to the HFI
- Driver sets up “send contexts” for user applications
- User applications get direct access through memory mapped registers
- Allows user application to achieve kernel bypass for the data path
- Utilizes a credit-based mechanism to coordinate filling buffers

Send Context from Kernel POV

- Abstraction for a PIO Send Engine
- Enables multiple ordered streams of packets
- Software representations of HW contexts



PIO Use Cases

- Best for small messages
  - Better for the CPU to move data
- Does cause contention on PCI bus
- No ordering between contexts or with SDMA Engines

Allocation of Send Contexts

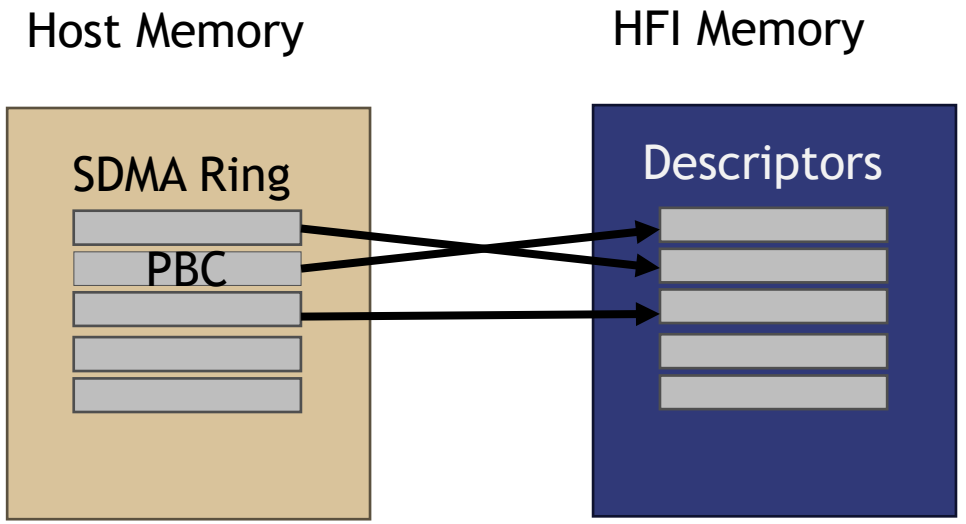
- Kernel keeps a small number of send contexts for itself
  - Need to be able to service VL15 and each data VL
    - Send context can serve multiple VLs
  - Verbs
  - IPoIB
- CPORT consumes a couple
- Most contexts are available for userspace
- User opens Cdev and gets a software context
  - Associated with this is a send context and receive context
- What about when there are more cores than contexts?
  - Valid concern as CPUs are getting bigger and bigger (in terms of core count)
  - Up to software how best to share those contexts

Filling Send Buffers

- Packets are ordered up to launch within a send context
- The send buffer is divided into 64byte send blocks
- Start of packet block begins with PBC
  - Describes the packet, tells the HW the length
- SW has to manage not overwriting unsent data
- HW “returns” credits
- User space has direct access to registers (kernel bypass)

SDMA

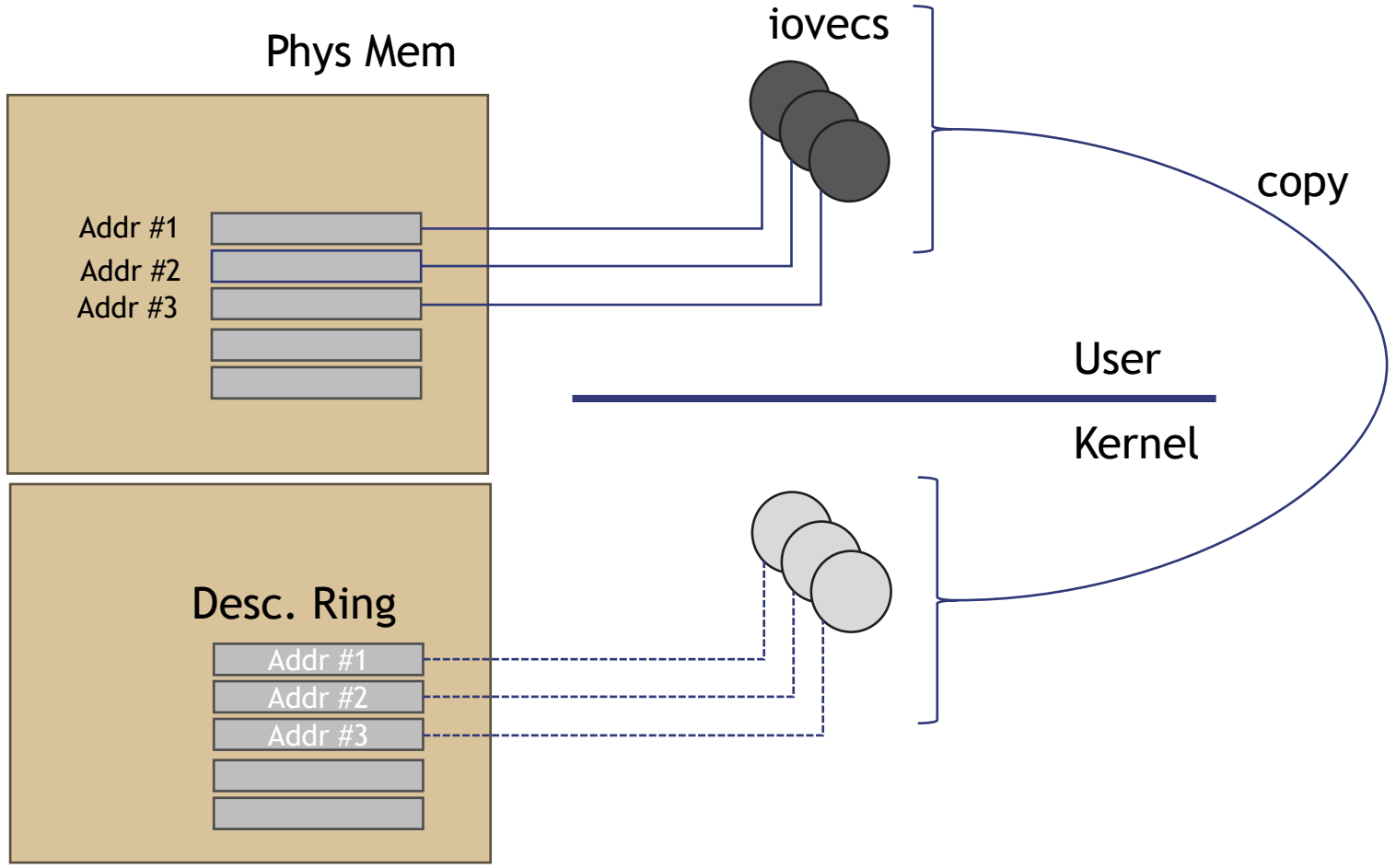
- Best for larger messages
- Requires setup
- Uses DMA engines on the HFI to move data
  - 16 engines
- DMA Engines programmed by the kernel
  - DMA Engines are shared resource
  - Only 16 of them



SDMA User/Kernel Interface

- User supplies list of data and header template
- Kernel does NOT copy data
  - Only pointers to the data
- Kernel builds packets (or HW)
  - Pins memory
  - Returns to user before send happens
- Kernel writes a “scoreboard”
  - Tells user the data has been sent

SDMA



Eager Receive

- Analogous to PIO Send but for Receive
- Driver setups of Receive Contexts
- User has direct access
- No architectural requirement that RC map to SC
    - Common use case though

Expected Receive

- User application registers TIDs with Kernel
- Kernel programs TIDs into the HW
- Data lands directly into user buffers described by TIDs
- KDETH Packets make this possible
- Rendezvous protocol commonly utilized
  - Ready to send
  - Clear to send

Mapping Packets To Receive Contexts

- Type 0 = Expected receive
  - TIDCtrl field is not 0
  - BTH DestQP maps to context
- Type 1 = Eager Receive
  - TIDCtrl field is 0
  - BTH DestQP maps to context
- Type 2 = IB Packet
  - BTH DestQP compared with QP Mapping table in HW
- Type 3 = Error
  - Default receive context
- Type 4 = Bypass (16B packets in WFR)
  - Preconfigured context is used

RSM Rules

- Think of this as an override for RX placement
  - For eager receives
  - Expected receives get reclassified as eager
- Existed in WFR (4 instances)
- Expanded in JKR (32 instances)
  - 1024 bits able to be inspected
  - RSM instances have 4 “match units”

Currently no user API

- What would you want to see?
- How could MVAPICH take advantage?
- This is where our dedication to Open Source shines
- User Interface is a hot button Kernel Topic
  - We need to collaborate

Being Micro Managed

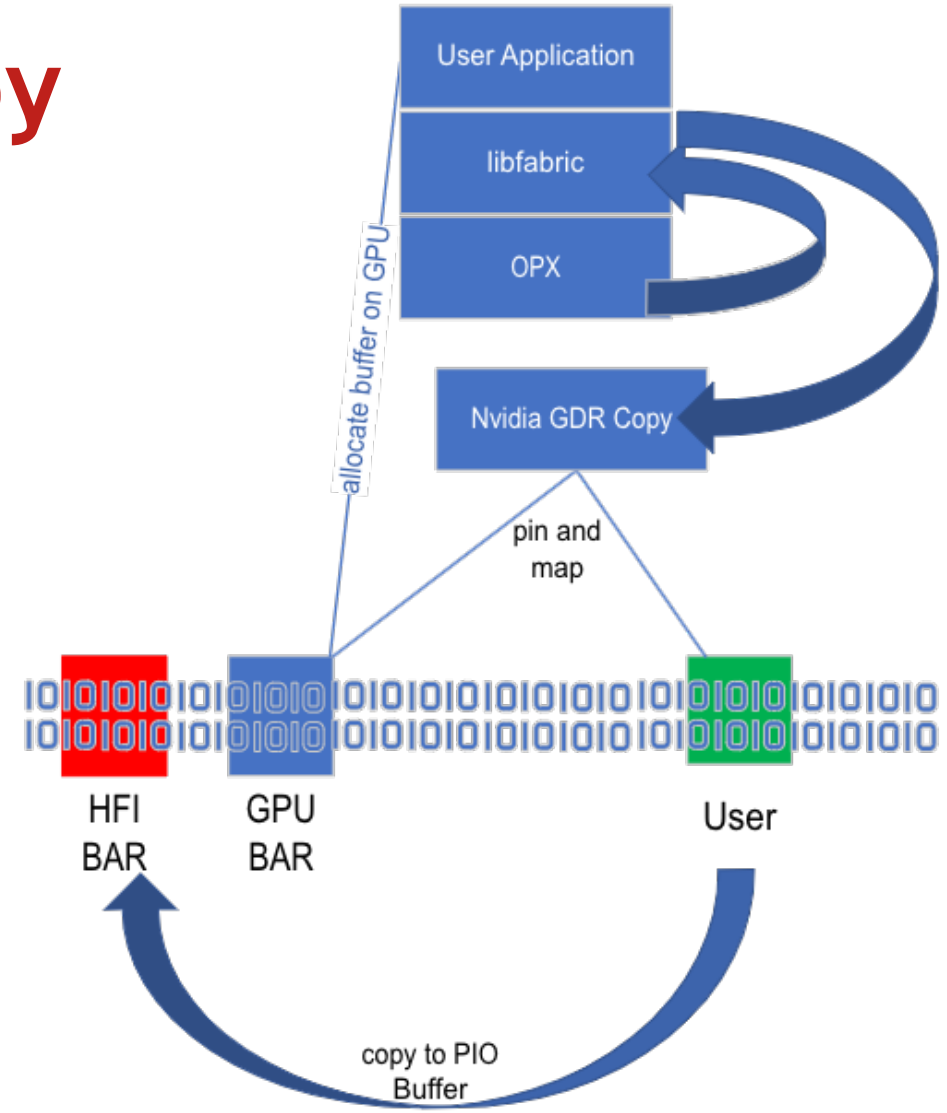
- New concept of tying in CPORT
- CPORT handles:
  - Firmware
  - Link negotiation
  - MAD packets
  - Telemetry
- Driver is free to do what it needs to do to support the user's job

GPU Support

- How does OPA take advantage of GPU Hardware?
- We will support Nvidia and AMD GPUs
- Required to use OPXS Version of the driver
  - GPU API Licensing and driver availability in Linux Kernel
  - DMA Buff only gets us halfway

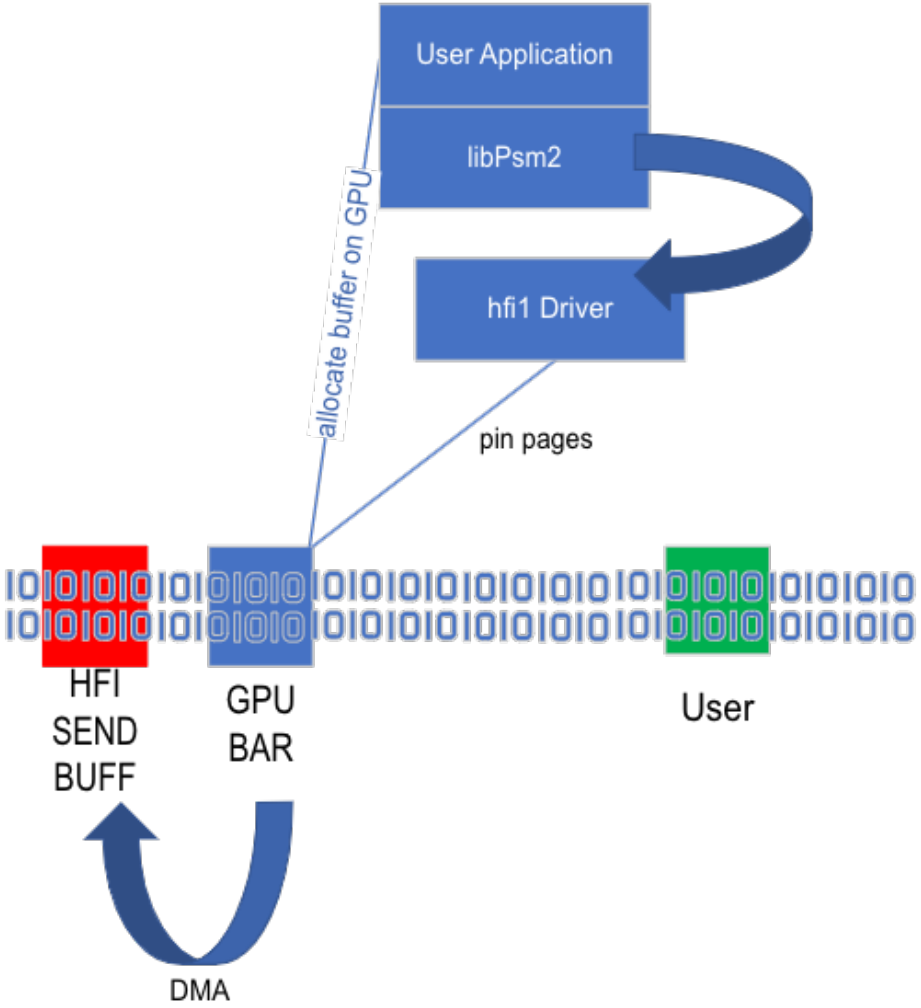
PIO Send – Device Copy

- GPU vendor provides device copy driver
- hfi1 driver is not involved
- Libfabric abstracts with hmem
- Just like any PIO Send



SDMA Send – Direct Access

- User app gets a GPU Buffer and tells hfi1 about it
- hfi1 pins GPU pages
- HFI DMA engines move the data



Part 3: Upstreaming

Working Together...Upstream Pathway

- OPX via Libfabric somewhat easier
  - We are active contributors and maintain our provider
- Kernel is a different story
  - Upstreaming major changes is not easy
  - Good thing we have been at this for a long time!

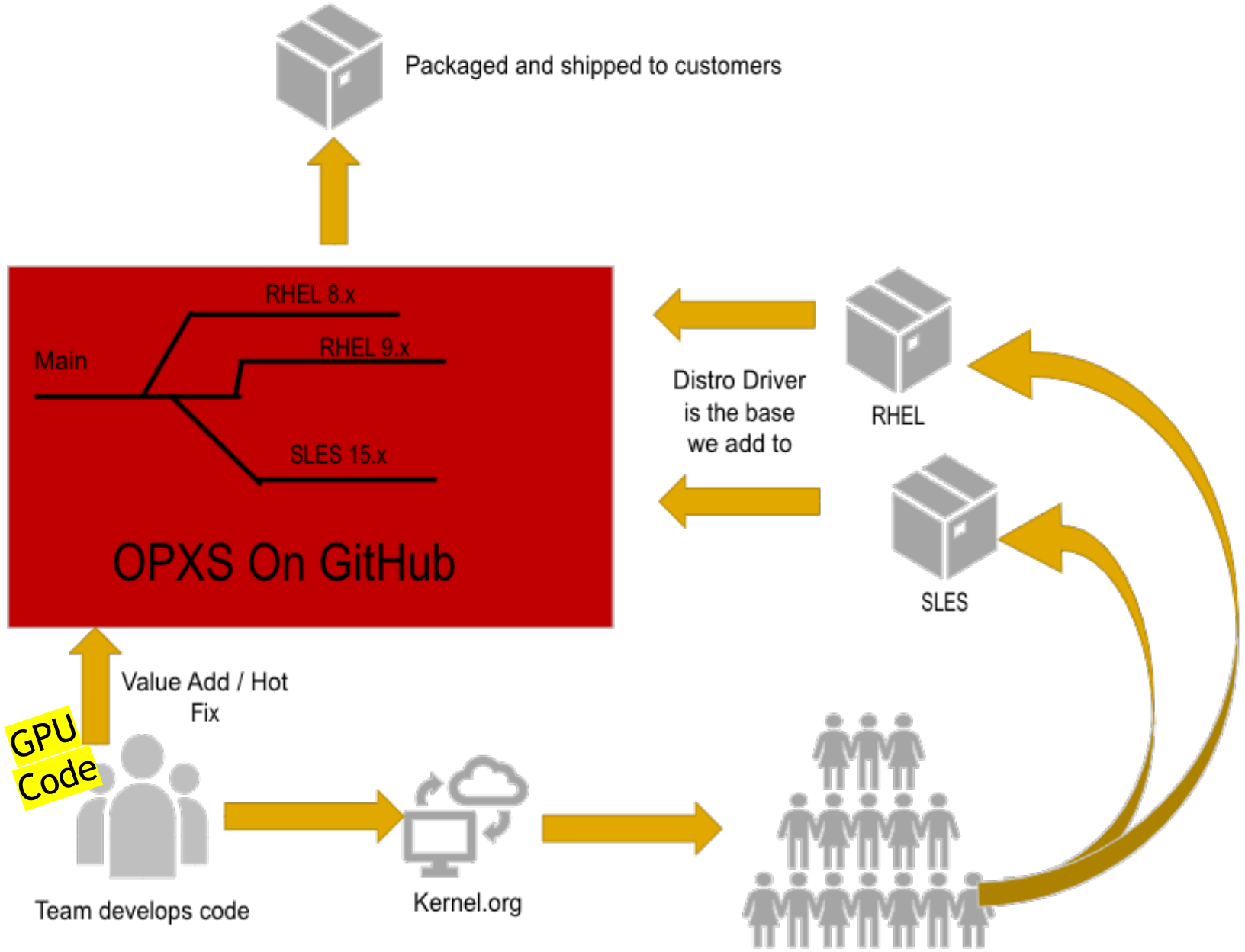
```
commit f48ad614c100783be1e7e777dc36328001b83999
Author: Dennis Dalessandro <dennis.dalessandro@in[redacted].com>
Date:   Thu May 19 05:26:51 2016 -0700

    IB/hfi1: Move driver out of staging

    The TODO list for the hfi1 driver was completed during 4.6. In addition
    other objections raised (which are far beyond what was in the TODO list)
    have been addressed as well. It is now time to remove the driver from
    staging and into the drivers/infiniband sub-tree.

    Reviewed-by: Jubin John <jubin.john@intel.com>
    Signed-off-by: Dennis Dalessandro <dennis.dalessandro@intel.com>
    Signed-off-by: Doug Ledford <dledford@redhat.com>
```

Old Email Do Not Use

CORNELIS NETWORKS

What is the Upstream Plan?

- Is this going to be hfi2?
  - One driver to rule them all ... and in the kernel bind them
  - JKR is based on a lot of the same concepts as WFR
  - Bigger, faster, better, new bells and whistles
  - Do we really need the 1?
    - Not really but why bother chaning
- Plan is to finally delete qib
  - Few known users of qib left in the wild
    - They keep popping out of the woodwork periodically though!
  - Product has long been End of Life
  - Delete qib as part of JKR upstreaming



Major Changes to hfi1

- Support multiple chips with a single driver
- Register changes and moves
- Support more than 1 fabric port
- Adapt to CPORT mgmt model
- uAPI changes (leverage RDMA core cdevs)
- Even more GPU support

Rdmavt

- Software verbs implementation
- Solved code duplication between hfi1 and qib
- With qib gone and hfi1 supporting both JKR and WFR do we still need rdmavt?
  - Technically, NO
  - However, no plans to remove it and collapse back into hfi1
  - Maybe someday if we run out of other things to do
  - Invisible to application writers

What's our overall status?

- Kernel code is coming ... Soon
  - watch linux-rdma
- CN5000 HW coming later this year!
- We still have OPA-100 too!
  - Lots of users
  - Still fully supported
  - Active development
    - Recently added AMD GPU
    - Recently added a backwards compatibility shim for libpsm2 cuda

Thank You

www.cornelisnetworks.com

CORNELIS™ NETWORKS