2024 OFA Virtual Workshop

Cornelis Networks CN5000 Adapter and Software Update

Dennis Dalessandro, Kernel SW Dev Manager

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.

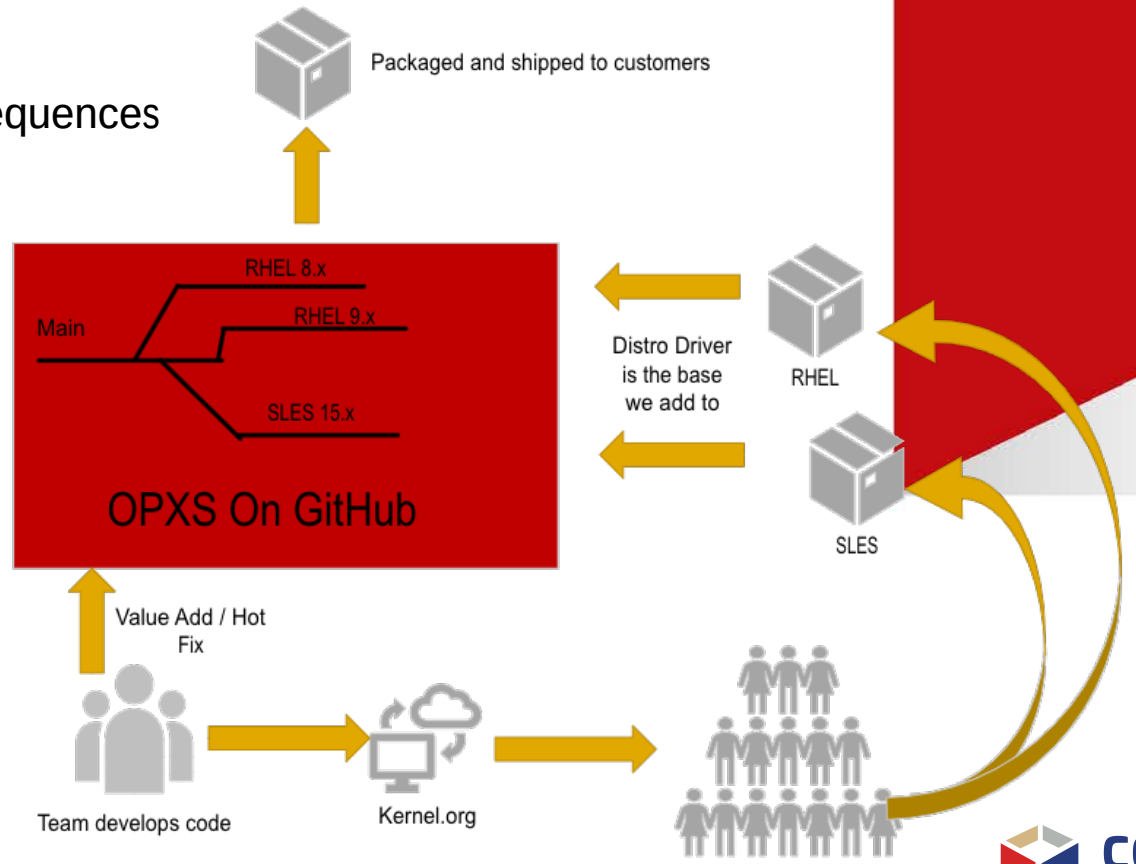
Cornelis Networks technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration.

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

Last Year

- Supporting an Upstream First Kernel Driver for HPC Fabrics
  - We covered the Why and the How
  - We talked about what not to do and the consequences



This Year

- How do you do upstream first with new and in development HW?
  - We are figuring that out!
  - What are we actually doing?
  - What is this CN5000 thing?
  - Challenges
- Not A Sales Pitch!
  - We got people for that, I can put you in touch if you want
  - The technology should stand for itself
- **My goal: Get the code upstream for when HW is available**

What is CN5000?

- CN5000 is next generation fabric solution
- Consists of adapters, switches, cables, and software
- We will focus on adapter and software for it
- Why?
  - That's what the host sees
  - SW is what people interact with
  - (frankly it's what I know!)

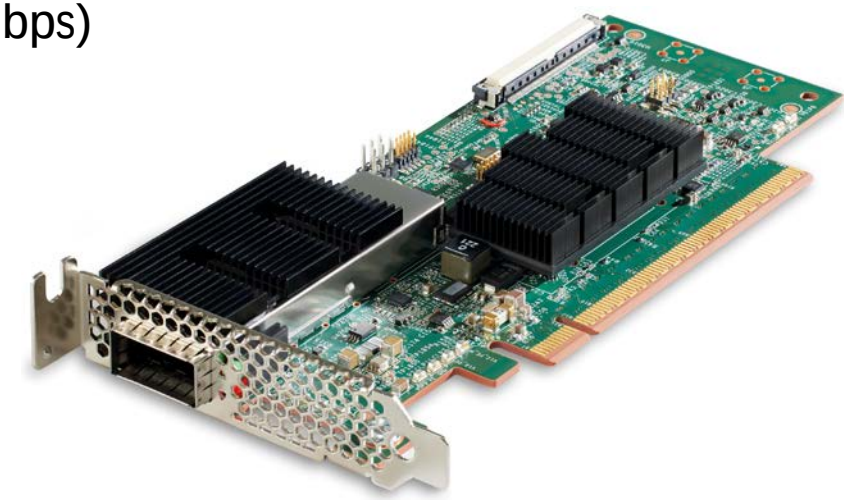
Switches

- 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)



Technical Details

- OPA-100 adapter ASIC is known as Wolf River (WFR)
  - Name leaked from Intel long ago – old news
- CN5000 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

Why am I telling you this?

- Shouldn't this be secret? --- Simply put, NO
- Everything I mention here is or will be obvious in the code
- Makes Open Source acceptance of the code easier… maybe?
- Open Source community can contribute more readily
- Settings/tweaks available in our tuning guide
- Honestly it's just really cool stuff nerds like to hear about
  - If you are bored now, just wait till we talk about code next!

Where were we? More details…

- 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!

What is the Upstream Plan?

- Is this going to be hfi2?
  - Discussed with maintainer last time OFA was in person response was: “Please God NO!”
  - I agree so it will be hfi1 still
  - JKR is based on a lot of the same concepts as WFR
  - Do we really need the 1?
    - Not really but why bother
- Plan is to 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

Rdmavt

- Software verbs implementation
- Presented to OFA a number of years ago
- 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

Major Changes: Registers moving

- While JKR and WFR are sort of similar a large number of HW registers have changed
- A large number of patches that “parameterize” something
- Then a follow-on patch to use that parameter
  - Might squash, we’ll see
- Added some new header files to separate WFR and JKR registers
- Not taking the qib approach of function pointers for each chip

```
commit b54fe09adca5457178320bf95ec46175164e6285
Author: Dean Luick <dean.luick@cornelisnetworks.com>
Date:   Thu Oct 19 10:45:01 2023 -0400

    RDMA/hfi1: Parameterize PIO init register

    The PIO init register will move in the new hardware.  Make it
    a parameter.

    Signed-off-by: Dean Luick <dean.luick@cornelisnetworks.com>
```

```
commit 6e8aa91b07f0ccbc86f33924062cb87f7983f3cb
Author: Dean Luick <dean.luick@cornelisnetworks.com>
Date:   Thu Oct 19 12:45:17 2023 -0400

    RDMA/hfi1: Add JKR PIO init register support

    Add the JKR PIO init register.

    Signed-off-by: Dean Luick <dean.luick@cornelisnetworks.com>
```

Major Changes: Multiple Ports

- We tried our best in WFR to be extensible to multiple ports
  - Number of places we decided to use per_port_data[0] blindly
  - Need to propagate the port number through multiple layers of code
  - Not technically difficult just tedious and a lot of code churn
  - 7 years of code being moved around and modified
- hfi1 has a per device data structure the ‘dd’
- It also has a per port data structure even for the single port case the ‘ppd’
- Moved a number of fields from the ‘dd’ into the ‘ppd’

Major Changes: CPORT

- WFR had extensive handling for MADs
- MAD processing is moving to CPORT
  - Enables more fabric security
  - Takes burden off of the driver, free to handle other packets
- Driver and CPORT exchange information through a register interface
  - Can also talk over the loopback ports for highspeed communication
- CPORT manages things like Link Status and LIDs, and PKeys, etc.
  - Driver still needs to know when things change
  - Driver has to be able to hand CPORT MAD packets from umad

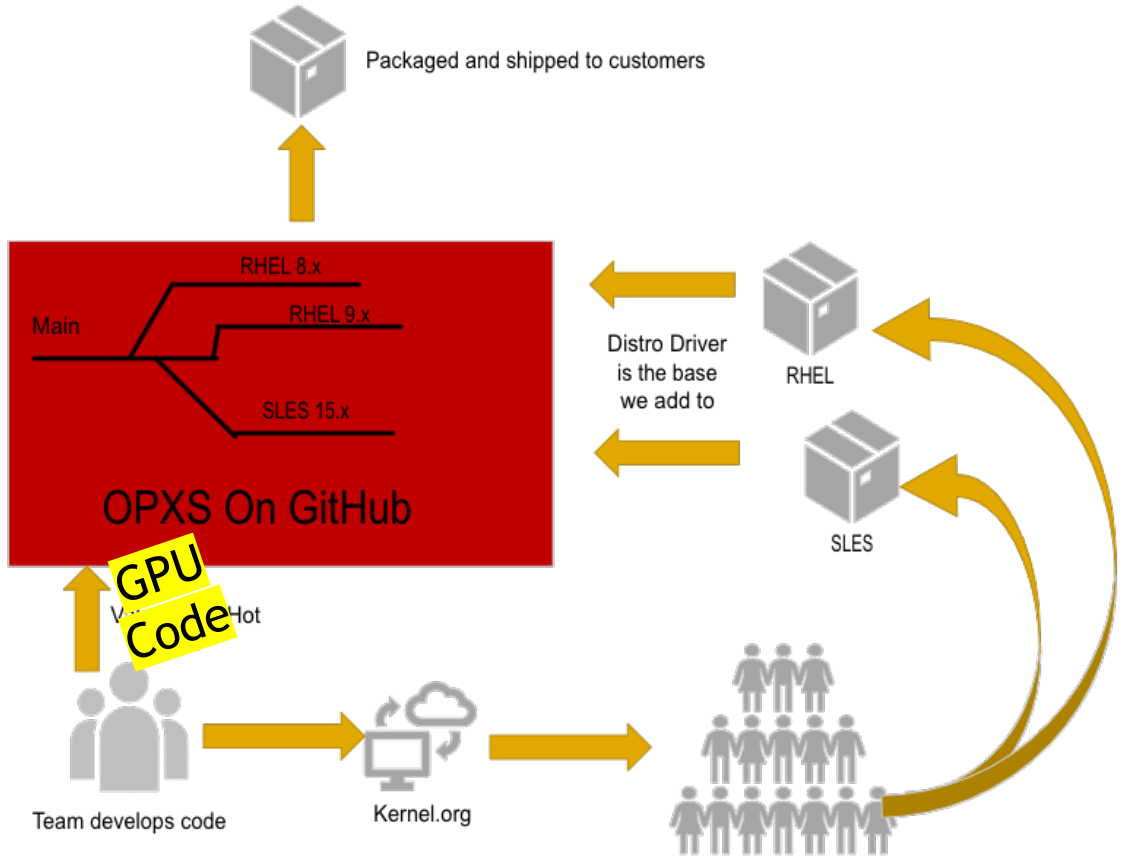
Major Changes: User API

- Technically no reason it has to change for JKR
  - Mostly: Needs to know what port to use, it does not currently
- We agreed long ago that the dreaded Cdev was on its way out
- Heavy upstream resistance to any changes in our user API until Cdev gone
- Likely plan:
  - Use uverbs FD for command and control
  - We still need an FD for data path
    - For SDMA submission, driver programs DMA engines
    - PIO is kernel bypass
  - Want to use io_uring()
    - Some complications, open issues
  - Need memreg for DMA Buff and GPUs
- Code is still a work in progress, look for an RFC soon

What about GPU?

- Applies to WFR and OPA-100 as well as JKR and CN5000
- Everyone knows the problem: GPU code not upstream
  - Not going to harp on this, I hope we can all agree its just bad
- What do we do as an upstream first development organization?
  - Other than hold our nose and just deal with it
  - Embrace the distro kernels and drivers in particular
  - Distro code + GPU = Our GitHub driver
  - Distro code comes from Kernel.org
  - Kernel.org provides no good way to hook in non-upstreamable code
    - Penalizes users in hopes they will push back on unfriendly-to-opensource vendors
    - Hasn't happened, not going to happen
    - There are technical issues too – how do you handle header files? License issues, etc.
    - Point is instead of being obstinate we as a kernel community COULD work the problem

Picture is the same for GPU



GPU *Motivated* Changes

- If we can't upstream GPU code we want to limit how much GPU code there is
- Turn out different GPUs and even our own ways of handling memory are similar
- Abstract out common code and "modularize" it
  - Move system memory page pinning to its own file and create a nice interface
  - Now we can "drop in" files for Nvidia and AMD that do their specific page pinning
  - Makes it easy to add/remove GPU code and limits difference between upstream
  - Have to monitor system pinning changes for upstream fixes and evaluate if needed in GPU

So what about DMA Buff?

- That’s a step in the right direction
- Only 1 GPU supports it, maybe 2 someday????
- This is part of our motivation to use verbs device
  - For memory registration, dmabuff support already exists
  - dmabuff will be a drop in addition like Nvidia and AMD code
    - Except it will be upstream!

Software High-level Status

- opa-fm
  - In-distro, will support CN5000 out of the box
- libpsm2
  - Active as long as OPA-100 is around
- OFI/OPX/libfabric
  - Available for OPA-100 but really targeted for CN5000
- Fast fabric tools
  - In-distro, some are being revamped, more details soon
- Kernel
  - Coming soon to a mailing list near you!

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