Modeling Composable Disaggregated Infrastructure

Presenter: Ryan E. Grant

Student Credits: Curtis Shorts, Jacob Wong, Shaina Smith, Jordan Abt

Collaborators: Yogi Boniface (DRDC), John Prokopenko (Cerio)

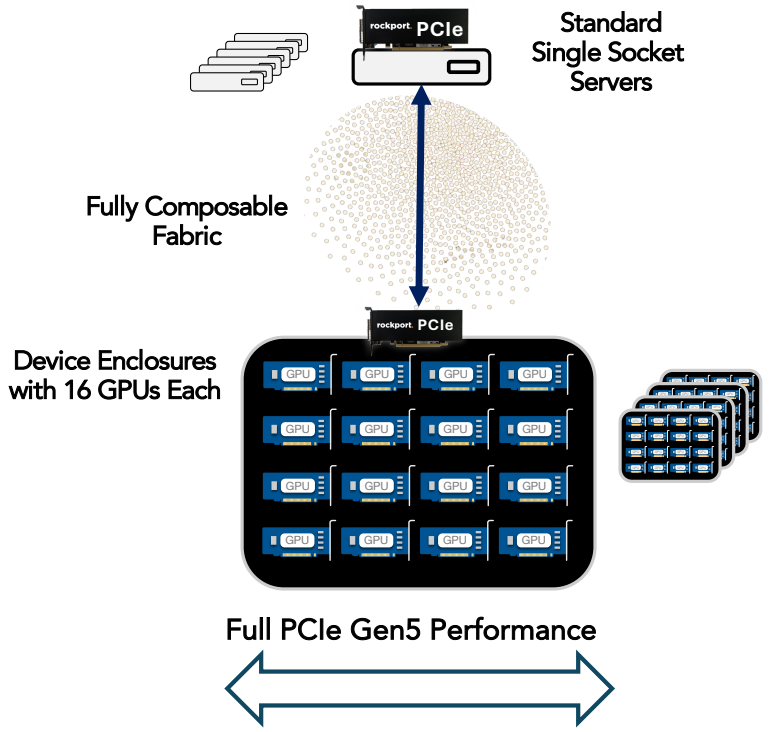
Outline

- Composable Disaggregated Infrastructure
  - Introduction
  - Modeling
  - Experiments
  - Modeling for field applications
- Heat Reuse
  - Modeling for infrastructure design

Disaggregated Systems and CDI

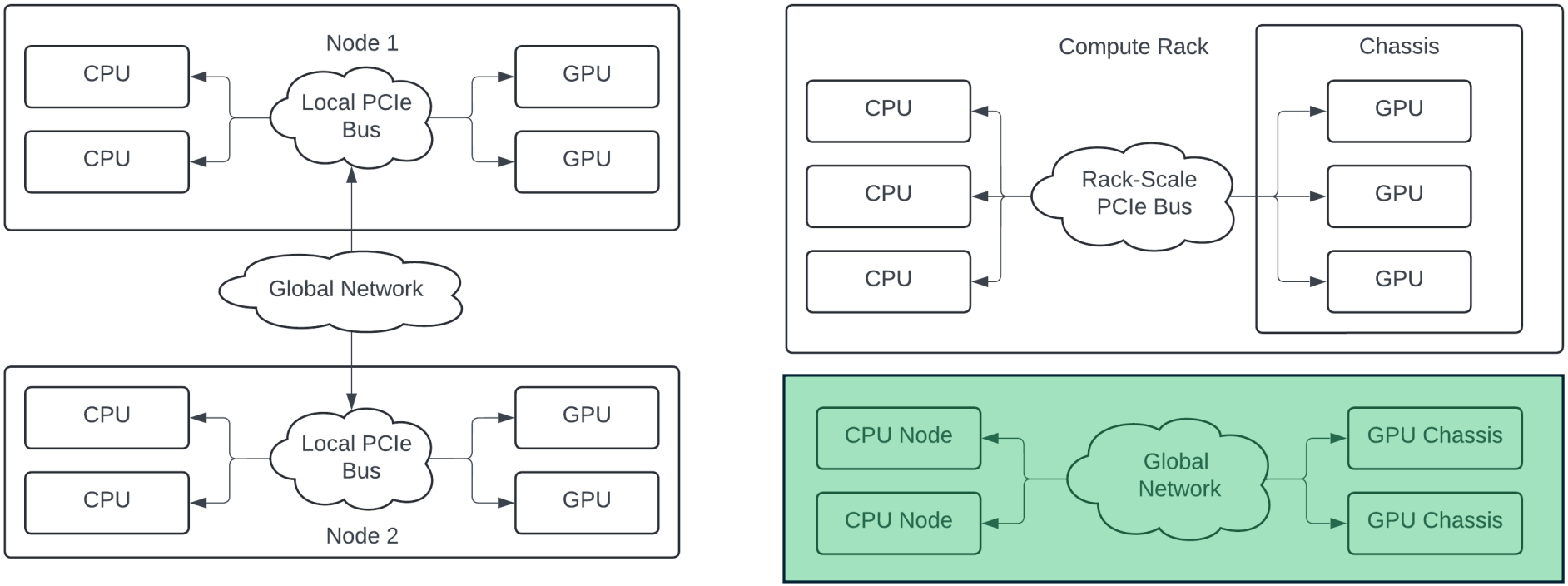
- CDI: Composable Disaggregated Infrastructure

- What is CDI?
  - Think of a system with CPUs and memory in traditional servers but with accelerators attached to specialized chassis connected to a CDI network.



Credit: Rockport Networks

Background – A Visual Representation of CDI

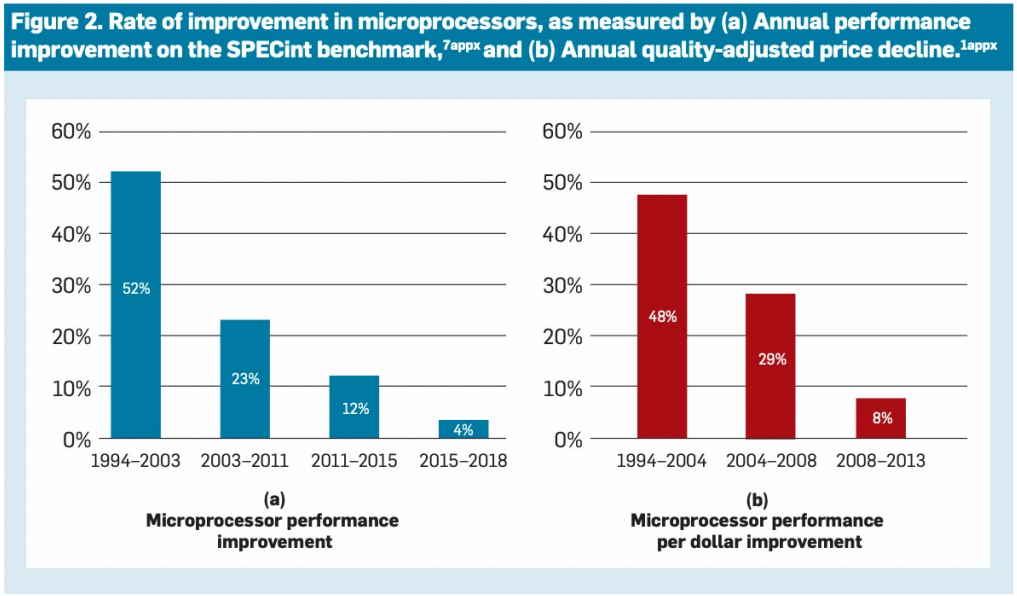


Visual examples of a traditional cluster architecture (left), rack-scaled CDI architecture (top right), and row-scaled CDI architecture (bottom right).

Why CDI?

- Need flexible compute
  - Not always the fastest but the most useful
- Optimize system setup for the task
- Many specialized chips for many purposes
  - Also, different classes of chips for compute

Why CDI?



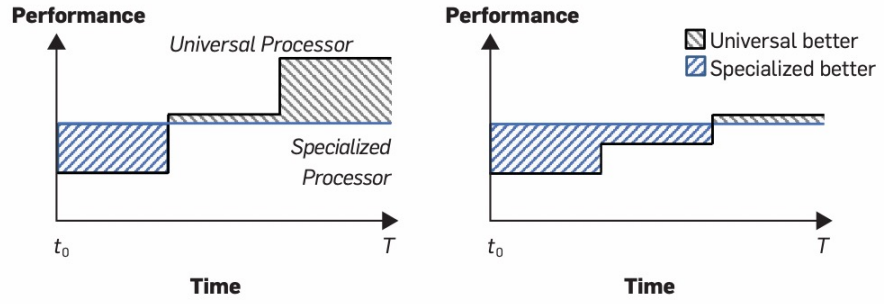
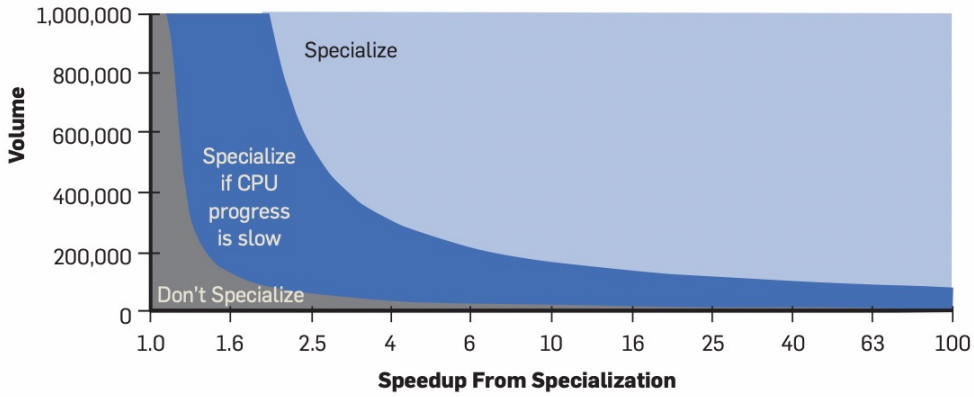


Illustration of processor choice trade-off when universal processor improvement rate is (a) fast or (b) slow



(c)
Model predictions for when specialization is preferred
CPU improvement rate per year
dark blue = 8% (slow), light blue = 48% (fast)

What Specialization means for CDI

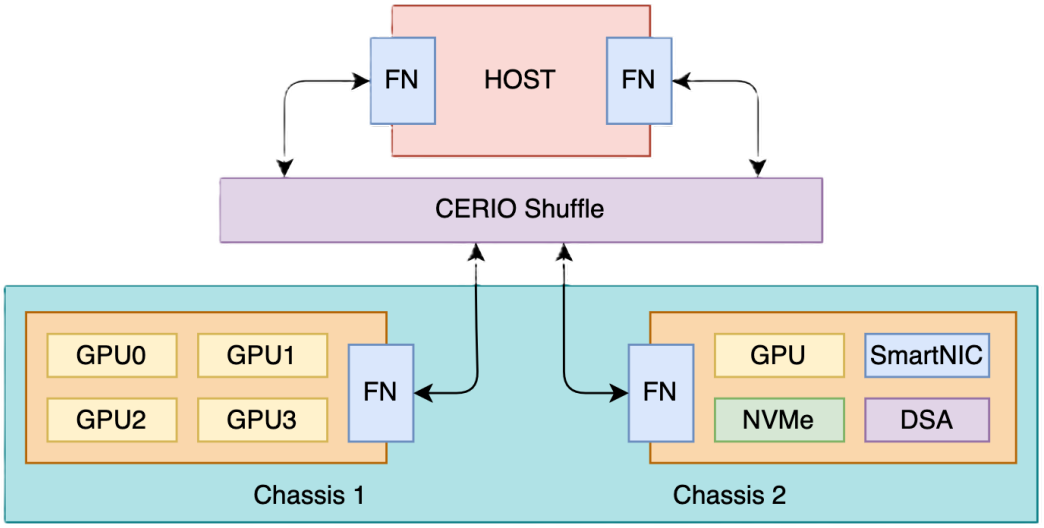
- Homogeneous system design is inflexible
- Many different kinds of specialized devices
  - Means we need many different node configurations
- But! Having many node configurations makes the system resource allocation inefficient
  - Why tie up a specialized accelerator (100%) that you only need for 20% of a job just because you need to have it for your job at one point?

CDI approaches

- Composable Disaggregated Infrastructure (CDI) - a heterogeneous HPC architecture that removes accelerators from the node and puts them in a resource pool (chassis) that nodes can compose devices from to make them look local
  - Decouples the scheduling of the CPU and GPU
  - Opens the door to interesting node CDI architecture
- Products such as GigaIO's FabreX achieve this with PCIe expansion technology
- Cerio's product uses a custom network to allow chassis to provision beyond the reach of PCIe

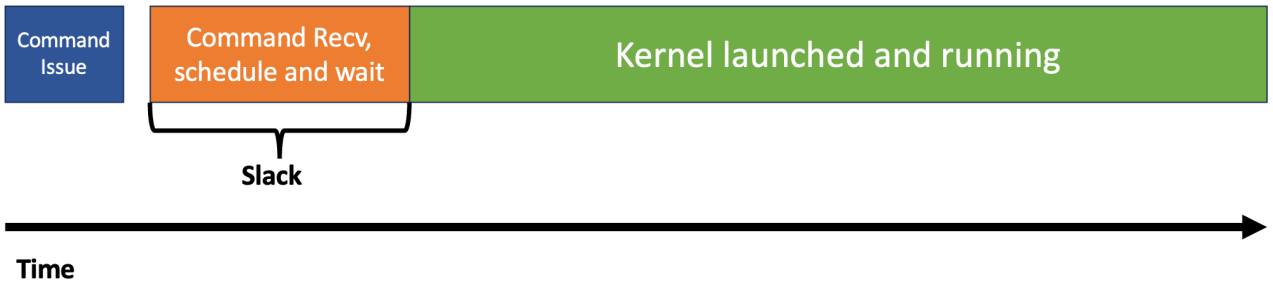
Modeling Cerio CDI

- Cerio CDI has a “fabric node” in each server and chassis (at least 1)
- Challenge: Determine the impact of FNs on large system deployments with different kinds of hardware



About those kernel queues

- The one challenge is keeping the GPUs busy with new work
- We've essentially moved the command issuer (CPU) further away from the workers (GPU)
- We're currently studying "slack" in command issuance and what we can get away with without hurting performance



But how much slack is there?

- Slack – the network delay introduced between the CPU and GPU due to the network
- If the GPU is supplied with sufficient work, latencies can be hidden through queuing and pipelined activities
- Extreme slack can expose latency by reducing the number of commands queued on the GPU

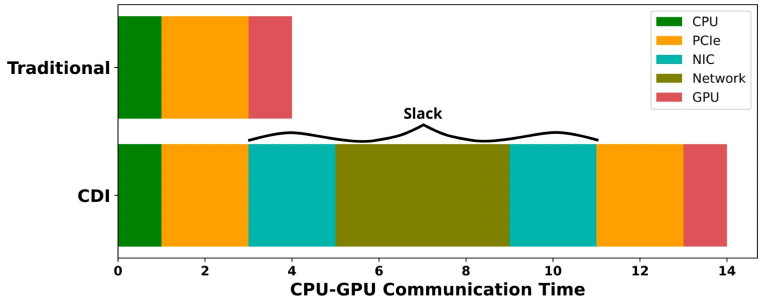


Fig. 1: An illustration of traditional and CDI CPU-to-GPU communication times. Slack is the time added by passing through the NICs and traversing network, as demonstrated by the annotation.

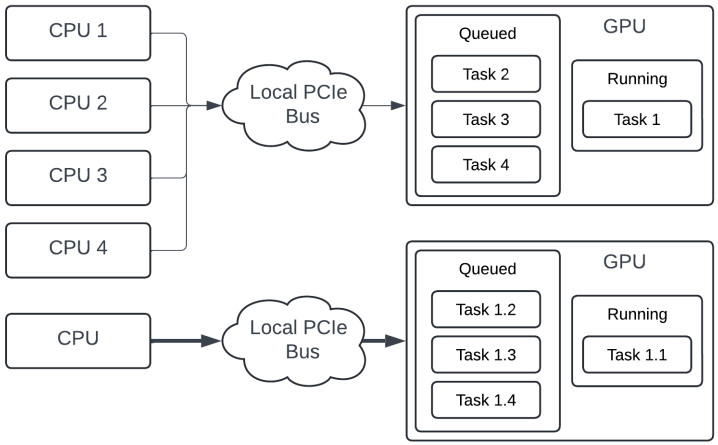


Figure 2 – A visual representation of "parallelization" in CPU-GPU queueing activities.

Modeling and verification

- Need to model GPU runtime variables:
  - Data transfer time, GPU kernel runtime, “degree of parallelization”
- Runtime degradation in proportion to slack is hard to understand and model

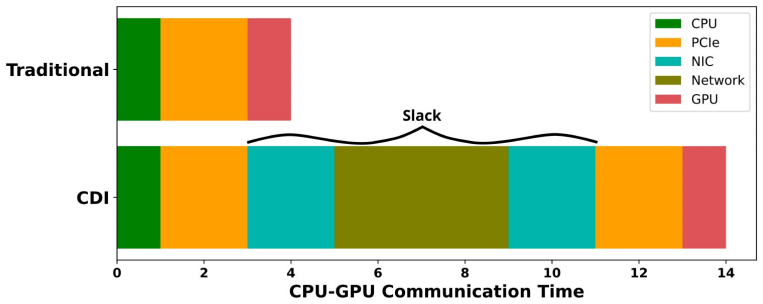


Fig. 1: An illustration of traditional and CDI CPU-to-GPU communication times. Slack is the time added by passing through the NICs and traversing network, as demonstrated by the annotation.

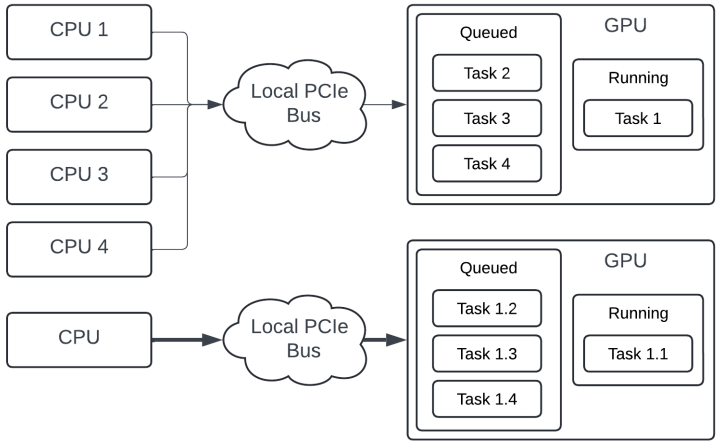


Figure 2 – A visual representation of “parallelization” in CPU-GPU queueing activities.

Slack and GPU scheduling

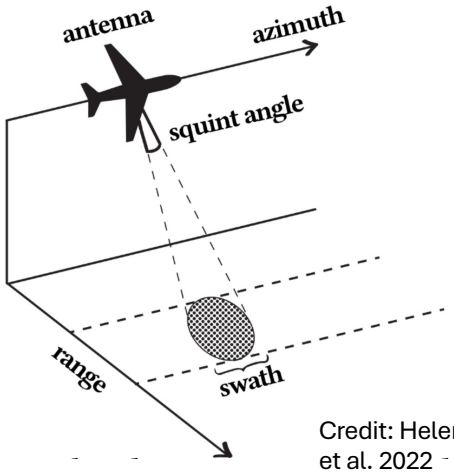
Research indicates that we have a fair amount of slack until it impacts performance significantly, 100us of slack is a long time in communication

	Slack [us]	1	10	100	1000	10000
LAMMPS Total Slack Penalty	Lower Value	1.000	1.008	1.009	1.025	1.417
	Upper Value	1.011	1.014	1.009	1.539	11.486
CosmoFlow Total Slack Penalty	Lower Value	1.002	1.002	1.004	2.871	24.501
	Upper Value	1.004	1.004	1.005	3.384	30.772

C. Shorts et al. “Examining the Viability of Row-Scale Disaggregation for Production Applications”, SC24 workshops

Modeling for a field CDI deployment

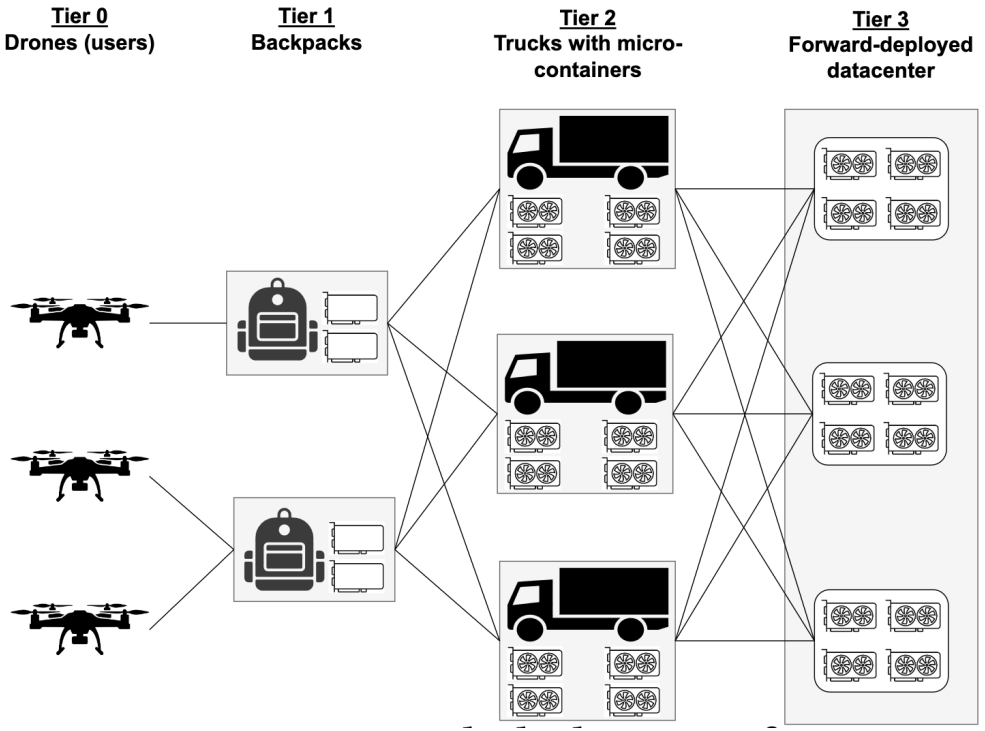
- Field application: Synthetic Aperture Radar
- Initial small-scale results show that maybe slack is manageable
- But large-scale deployment in field is impractical
- How do we model the large-scale field deployed
- Problem: Wildfires



Credit: Helena Cruz et al. 2022

CDI for SAR

- Can we use CDI to make field deployed devices look like they have a lot of compute power?
- Can we adapt the compute power as needed for situational needs?



But what magnitude is slack?

- We can get some approximate numbers with artificial slack insertion by intercepting CUDA calls
- Interesting that for high resolution images long slack times may be acceptable
- Long range communication is possible

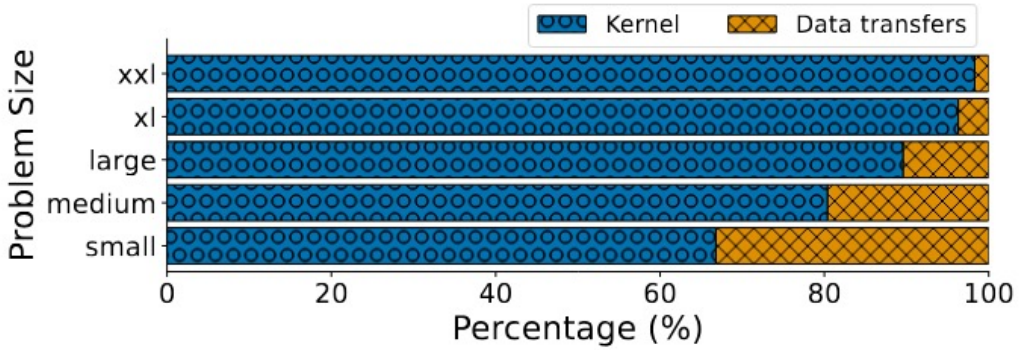
Data centre class GPUs (H100)

GPUs	Size	Base (s)	Slack 10μs	Slack 100μs	Slack 1 ms	Slack 10 ms
1	S	0.001	1.14	1.44	7.96	75.76
	M	0.007	1.02	1.04	1.58	11.48
	L	0.053	1.01	1.01	1.04	1.81
	XL	0.583	1.01	1.01	0.99	1.05
	XXL	5.077	1.00	1.00	1.00	1.00
2	S	0.001	1.55	3.44	23.63	224.90
	M	0.004	1.07	1.18	4.05	38.51
	L	0.028	1.01	1.02	1.21	5.70
	XL	0.282	1.02	0.98	1.01	1.35
	XXL	2.604	1.00	1.00	1.00	1.02
4	S	0.001	2.34	5.57	38.27	364.05
	M	0.004	1.09	1.32	8.80	83.76
	L	0.021	1.01	1.04	1.73	15.65
	XL	0.152	1.01	1.01	1.07	2.26
	XXL	1.386	1.00	1.00	1.00	1.07

But why?

- SAR is computationally expensive, as problem size grows compute dominates
- Means slack is even less of an issue for less capable GPUs

Compute grows faster than data transfer size



A40 passively cooled GPUs

GPUs	Size	Base (s)	Slack 10μs	Slack 100μs	Slack 1 ms	Slack 10 ms
1	S	0.027	1.01	1.01	1.14	2.66
	M	0.209	1.00	1.00	1.01	1.18
	L	1.650	1.00	1.00	1.00	1.02
	XL	12.058	1.00	1.00	1.00	1.00
	XXL	85.190	1.00	1.00	1.00	1.00
2	S	0.014	1.02	1.06	1.64	10.18
	M	0.106	1.00	1.01	1.07	1.84
	L	0.833	1.00	1.00	1.01	1.10
	XL	6.098	1.00	1.00	1.00	1.01
	XXL	42.960	1.00	1.00	1.00	1.00

Can optimize data transfers though

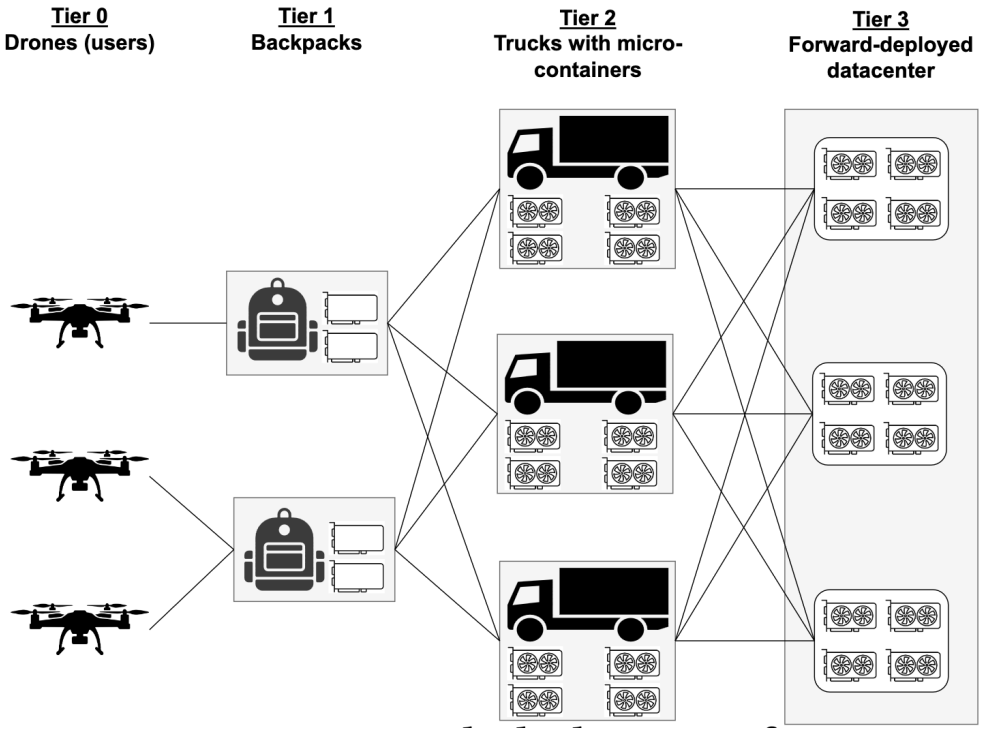
- Not a naïve transfer of all data to all GPUs in a chassis
- Send to a leader that then sends to others
  - Need all the data anyways
- Can help us optimize the overall transfer when we have limited bandwidth
  - Solves problems with field conditions that require lower bandwidth wireless solutions
- Built this into the model as well

Validation at small scale

- We then check the model predictions versus real transfer times with different sized problems
- Find 1-5% variance from real measurements at small scale
  - Always conservative – implies roofline but can't confirm
- We inject some local artificial slack by delaying GPU commands
- We can compare this at small scale to real results

Operational Modeling

- Remember the basic setup
- Lower power devices with fibre connections
- Different compute capacities at different ranges
- Also looked at wireless solutions at much lower bandwidth



Operational Modeling

- 2 scenarios, example 1 with ”short” fibre runs, example 2, longer
- Speedups versus a 2070 GPU equivalent
- Dark green is best solution, light is close to best and short distance

	Baseline	Example 1					Example 2			
Location	Drone	Drone	Drone	Backpack	Backpack	Datacenter	Drone	Truck	Truck	Datacenter
Distance	0 m	0 m	0 m	10 km	10 km	100 km	0 m	100 km	100 km	1000 km
Interconnect	-	-	-	10G Fiber	1G Air	1G Air	-	1G Air	1G Air	1G Air
Device	2070	ARM CPU	2070	2x A40	2x A40	4x H100	2070	4x A100	A100	4x H100
S	0.06	0.016	1	1.45	0.214	0.221	1	0.221	0.220	0.188
M	0.45	0.014	1	2.14	0.389	0.425	1	0.424	0.421	0.408
L	3.57	0.011	1	2.88	0.709	0.844	1	0.841	0.829	0.835
XL	26.74	0.008	1	3.54	1.21	1.64	1	1.61	1.49	1.63
XXL	188.58	0.007	1	3.89	1.77	2.85	1	2.78	2.46	2.85

Takeaways

- Modeling for CDI allows us to explore system deployments that would otherwise be impractical to test
- Verify that there is a high likelihood of success and improvement
- Delays from CDI are not as high as intuition would lead you to believe
- Simulation environments on their way
  - Working on understanding in more detail in the datacentre how the simulations work to match reality

Heat Reuse Modeling

Modeling Heat Reuse

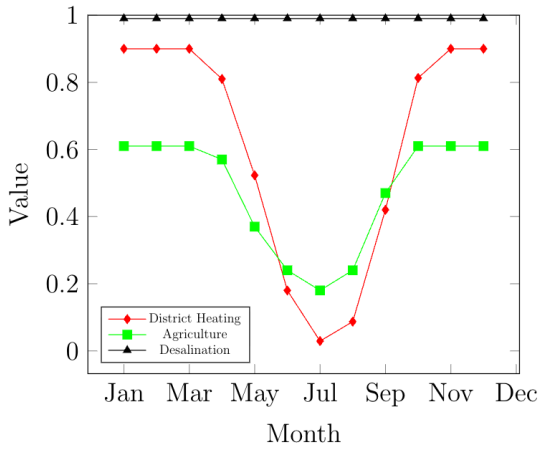
- Modeling and co-design for datacentre design
- Design with the idea that heat is not a waste product but has value

- How does this different approach change how you model?
  - Assume infrastructure that you wouldn't assume you need in a traditional build
  - Design for different running temperature points than you might otherwise
  - Think about water "upgrading" costs
  - Consider filtering more than a traditional centre

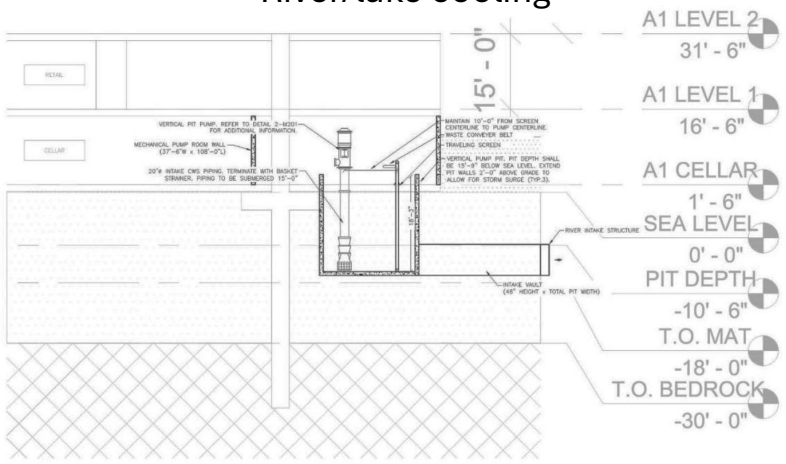
Heat reuse

- Thinking about heat use
- Also heat rejection
- Can work with our environment
- Can use hot water to make cold water
  - Adsorption/absorption chillers
- New metrics for green supercomputing

Heat recovery proportion



River/lake cooling



Build economic models from here

- Help to co-design building infrastructure and community infrastructure for everyone's benefit
- Exposes new purposes that weren't thought about before
  - Supercomputers and breweries?
- Economics doesn't drive the whole design
  - But it is considered as a first-class parameter

Takeaways

- Modeling helping to drive use cases
  - Discover new ones too
- HPC impact outside of the supercomputer
  - More use cases in the field
  - Community-minded design
- Rapid idea to model to small scale implementation
  - Accelerating with AI solutions

Thank You!

Questions?

DRDC | RDDC

Natural Sciences and Engineering Research Council of Canada

Conseil de recherches en sciences naturelles et en génie du Canada

Canada

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).