Putting Consistency, Reliability, Availability and Partition-tolerance in the SmartNIC

Paul Borrill, Jonathan Gorard

Plus: *Charlie, Steve, Liane, Chuck, Melissa, Susan*

SmartNICS Session: Wednesday, June 14th 04:20-5:20 PM

DÆ

DÆDÆLUS

Daedaelus?

- DÆDÆLUS addresses fundamental problems in distributed systems using protocols, data structures and algorithms inspired by Quantum Information Theory and Multiway Systems
- Our market is next generation platforms for secure, reliable, distributed computing on the edge
- We provide Microdatacenters with a fundamentally more reliable and programmable graph infrastructure
- Initial use-cases include Transaction systems, Digital Twins, AI/ML/LLM infrastructure, Multiplayer Games and Interfaces to Quantum Computers

See session: Wednesday, 04:20-5:20 PM

At the beginning of time, in networking

- A set of brilliant decisions were made
  - Packets could be dropped
    - For congestion
    - And to simplify handling certain corner cases
- TCP sessions
    - Would recover those packet drops, deliver in order
    - If the TCP session disconnected, recover in the app
    - Even re-running a file transfer over Arpanet wasn't hard

As the Hyperscale era began

- A set of brilliant decisions were made, again:
  - Commodity hardware and software, only
    - White box servers
    - Linux
    - NICs, and later switches
      - Existing, proven foundation technology only
  - Scale out, not scale up
    - Distributed databases
    - Distributed storage systems
    - Load balancers to replicated front ends
      - To stateful back ends

Distributed Applications are Hard

- Nodes need to agree on a lot of things, all the time
  - What nodes are in the cluster? Who's up? Is the "leader" alive?
  - Did that storage write (or database update) commit?
- Getting consensus algorithms right is hard
  - Lost packets, broken TCP connections: big impact
  - Gray failures (performance collapse) happen often (ZOOKEEPER-1465)
- What happens in a network partition is harder
- Partial network partitions are worse

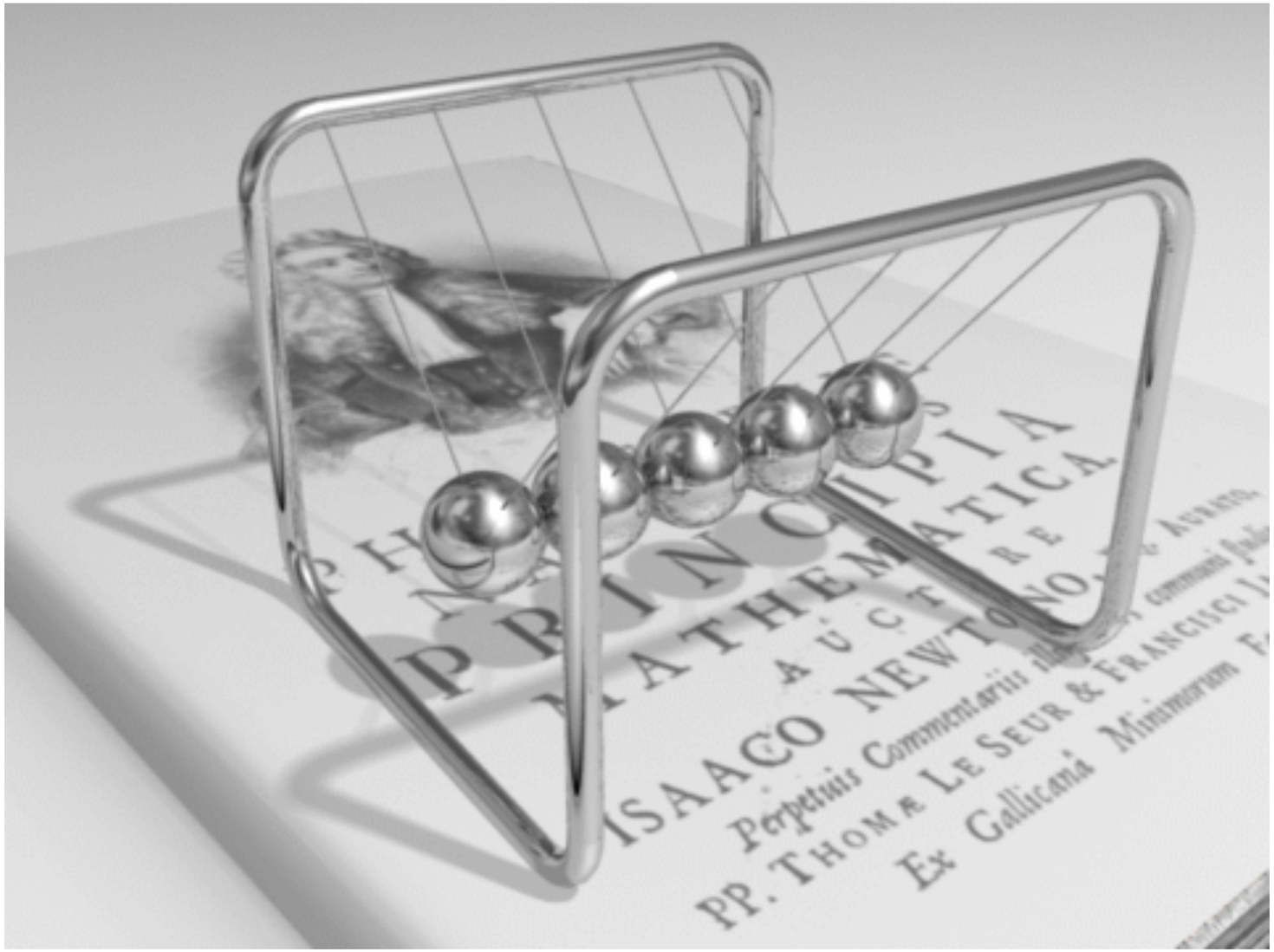
A Surprising (and Scary) Conclusion

- Brilliant networking decisions and brilliant hyperscale decisions together cause metastable failures in stateful applications like distributed storage and databases
  - At best cause performance collapse ("gray failure")
  - At worst cause silent data corruption (University of Waterloo)
  - Google SRE Handbook, chapter 23 "Managing Critical State"
- Computer Science has studied this at length, and concluded that these problems can be mitigated but not solved [Not true by the way]
- SmartNICs are in an ideal position to do *more* than mitigate these problems

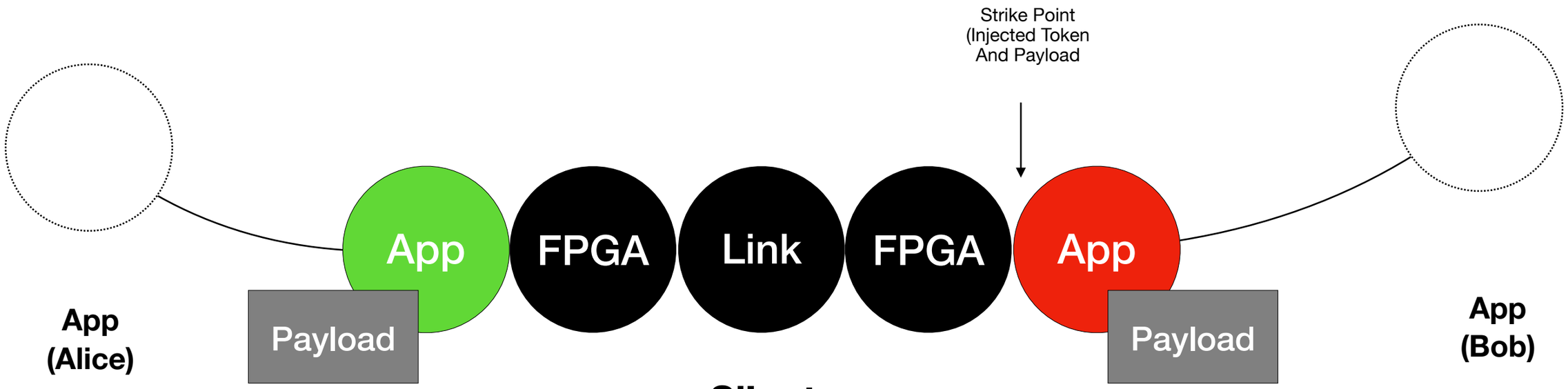
Can't solve this within "business as usual"

- We tried giving distributed applications reliable, deterministic communication
  - In a software layer over best practice networking
    - *The industry failed*
  - Using custom protocols and drivers over off the shelf NICs, with and then without switches
    - *The industry failed*
- We found the need for "entanglement" at the Link layer and end-to-end, so the sender and receiver both know *immediately* if a packet arrived successfully, without timeouts and retries

Newton's ~~Cradle~~ Cable



No Cloning Link Protocol



Silent
(cannot be observed
Don't even try)

Extraordinary claims require Extraordinary Proof

See our Table Top Demo in the Exhibit Hall

The Network Changes, But Doesn't

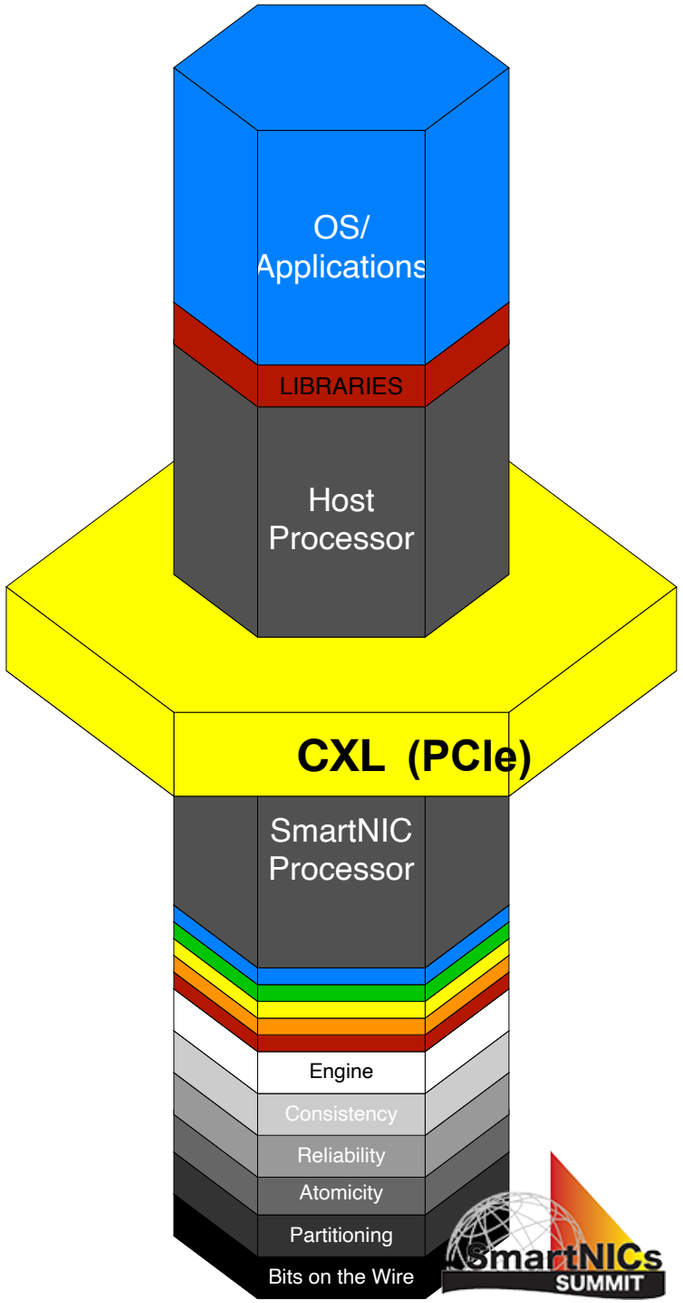
NICs connect directly to each other – no switches necessary

- Uplinks from the network are backwards compatible
- TCP/IP and Ethernet stack are unchanged

There are no dropped packets

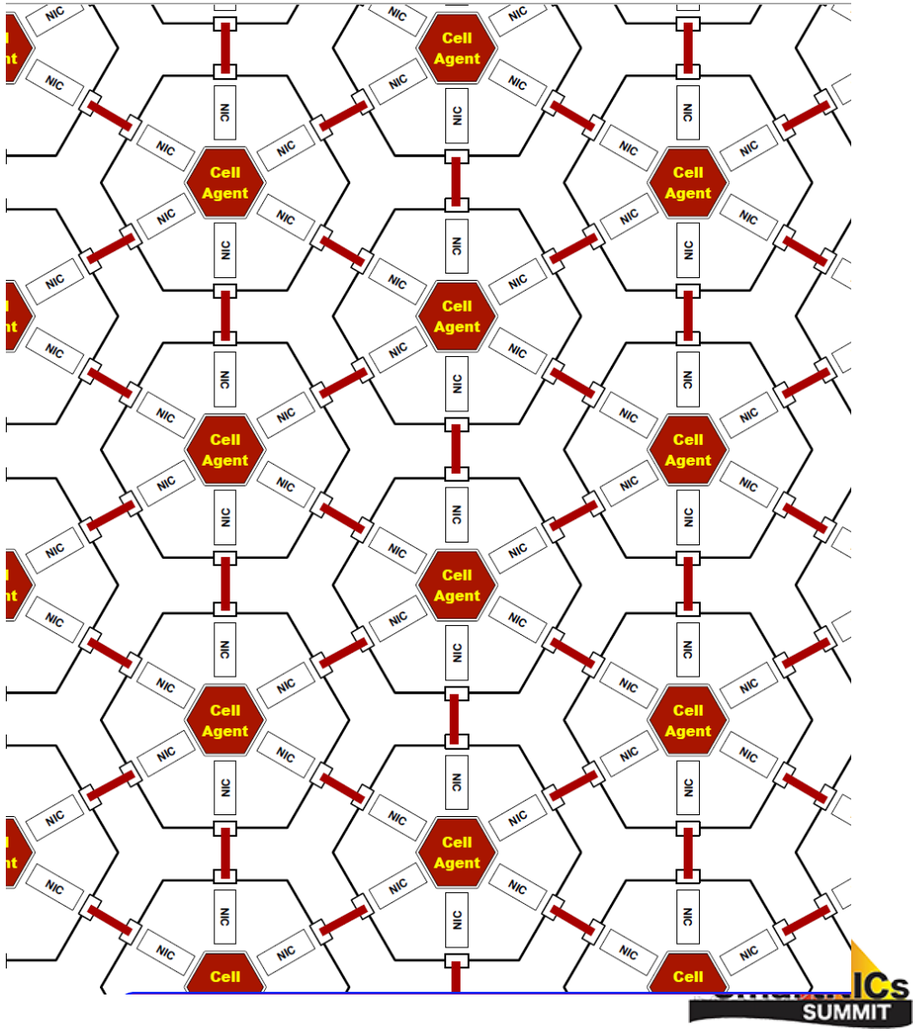
- Traffic is paused on link failure & healing – locally
- If a transaction packet doesn't reach destination, we know in microseconds
  - We don't use timeouts, causality/events on multiple paths
  - We ensure both ends have the same facts about whether a packet was delivered or not

At the application level: enables agreement on facts across a set of nodes, despite the CAP theorem "proof" can't be done



The Rack/Row Deterministic Subnet

- Uplinks from that subnet are just Layer 3 via Ethernet
- Addressing and forwarding have novel properties
    - Software endpoints are addressed, not servers
    - A software endpoint can move within the subnet
    - Endpoints can be managed in directed graphs
- New hardware paradigms are enabled
    - Server really is a peripheral of the NIC
    - Enables 10x more 10x smaller servers
        - 2 centimeter cables between adjacent nodes
        - Connection cost per server radically lower
        - Consensus which actually works enables dividing a distributed app over far more nodes
        - Endpoints in sets/graphs simplify management, deployment, and ACLs



Daedaelus

A graph ***software*** company focused on dependable computing

We solve putatively unsolvable problems in the communication between pieces of a distributed application

- Which reside on different computers
- Which communicate over a fallible network
- Which require agreement on certain facts in order to operate correctly

Incidental to our solution, we write code for an FPGA NIC

Incidental to our solution, we use a mesh network of servers

As part of our solution, our FPGA NIC provides line-speed extreme low latency primitives which assist consensus, atomic update of shared data items, conservation of tokens, etc for distributed app nodes within our subnet

Distributed Systems APIs for SmartNICs

- GREY FAILURE NEEDS OUR HELP
- THE CAP THEOREM NEEDS OUR HELP

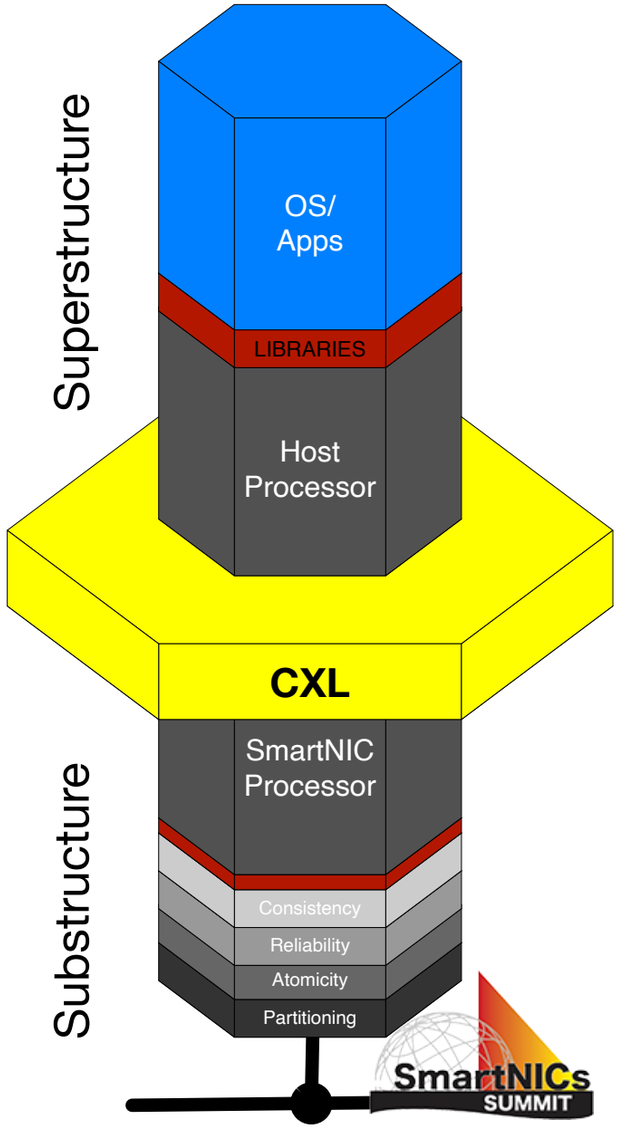
- KUBERNETES NEEDS OUR HELP
- PROGRAM WITH LINGUA FRANCA
- SECURE ENCLAVES NEED OUR HELP
- TRANSACTIONS NEED OUR HELP

IN HONOR OF MARK CARLSON

We make Transactions Reliable

- **Reliable Substructure Clusters**
  - Reversible transfers at line rates, with ultra-low Latency
- No tradeoff of Bandwidth and Latency
- No Metastable Failures

- **FPGAs do not have a halt state!**
  - They just run. Just circuits. Stuff goes in comes out, no halt states in between
  - Unlike ASICs, they don't take years to create

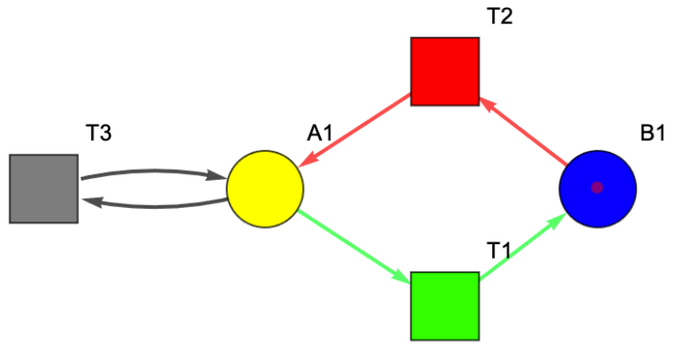


Labyrinth - Formal Verification & Simulation

1. Compile directly from Rust into the Dædælus protocol description language
2. Simulate all possible non-deterministic evolution histories with state transition graphs
3. Extract *entanglement* information from the protocols, indicating which microstates are non-separable (as in quantum mechanics)
4. Simulate typical failure scenarios (e.g. link failure, packet loss) and quantify robustness and recovery capability

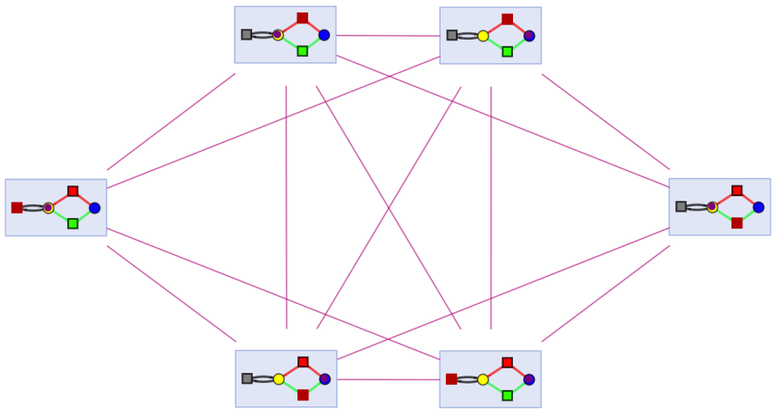
```
In[483]:= DaedaelusFullCompile["linkLiveness.rs"]
Out[483]= DaedaelusProtocol[ Places: 1+1  Transitions: 1+1+1  Arcs: 2+2+2  Tokens: 0+1 ]
In[484]:= %["LabeledGraph"]
```

Out[484]=




```
In[490]:= SimulateDaedaelusProtocolEntanglement[DaedaelusProtocol[ Places: 1+1  Transitions: 1+1+1  Arcs: 2+2+2  Tokens: 0+1 ], 2]
```

Out[490]=



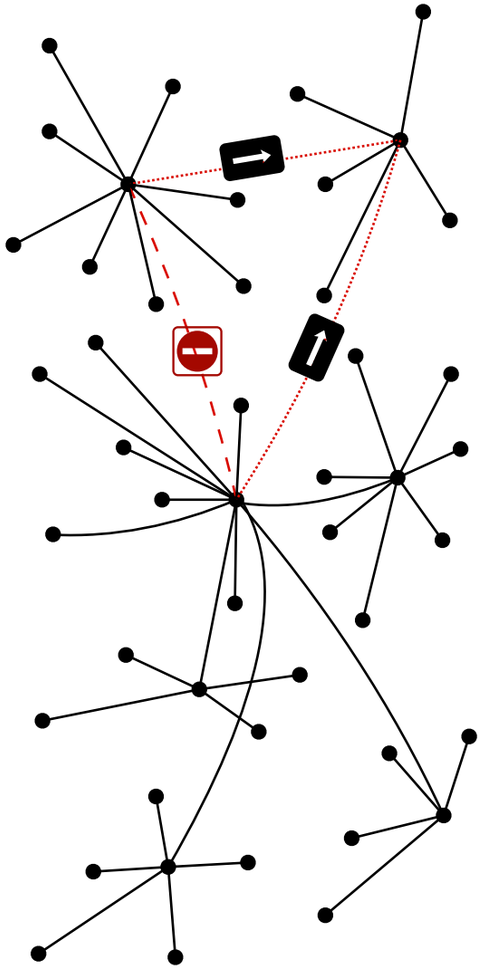
https://github.com/JonathanGorard/Labyrinth

Questions?

Paul Borrill
Founder/CEO and Team
info@daedaelus.com

DÆDÆLUS

Application errors caused by communication issues



- 80% of failures have a **catastrophic impact**, with data loss being the most common (27%)
- 90% of the **failures are silent**, the rest produce warnings that are unclear
- 21% of the failures lead to permanent damage to the system.
- This **damage persists** even after the network partition heals

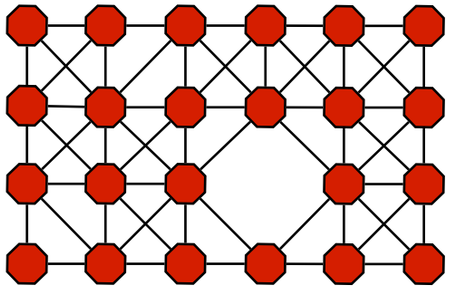
https://dl.acm.org/doi/10.1145/3576192

ACM Trans. Comput. Syst.

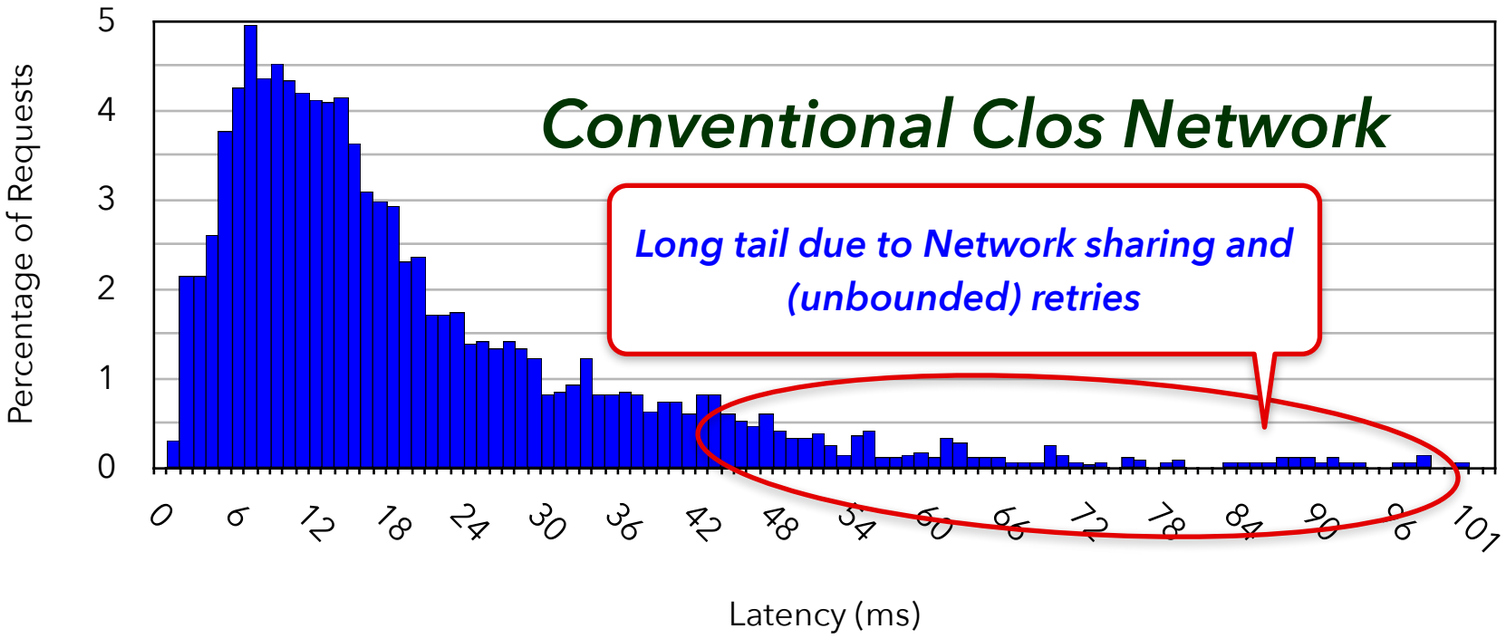
Tail Latency

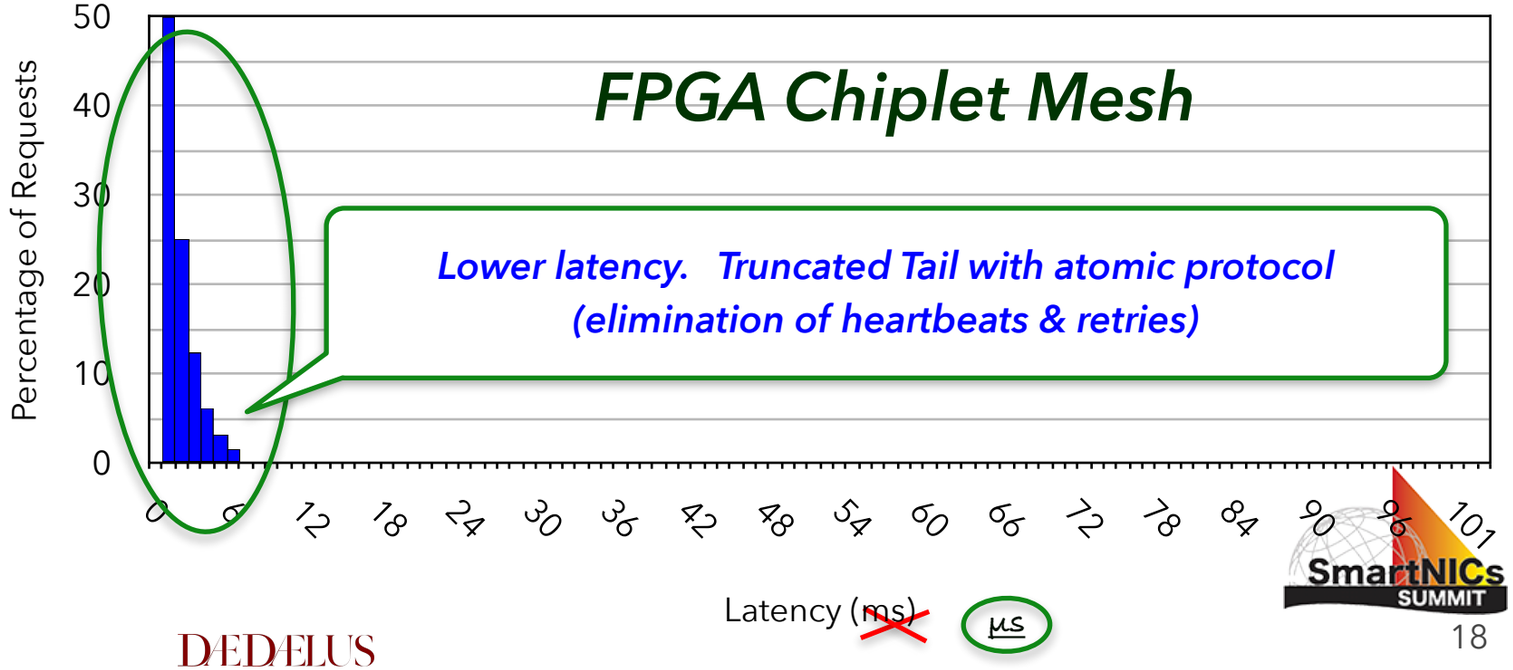
Daedaelus reduces latency in ways conventional networks cannot:

- Direct connections
- Multicast consensus, in parallel over 8 ports instead of serial over 1
- Truncated Tail Latency – protocol knows it failed or succeeded (without heartbeats or timeouts)

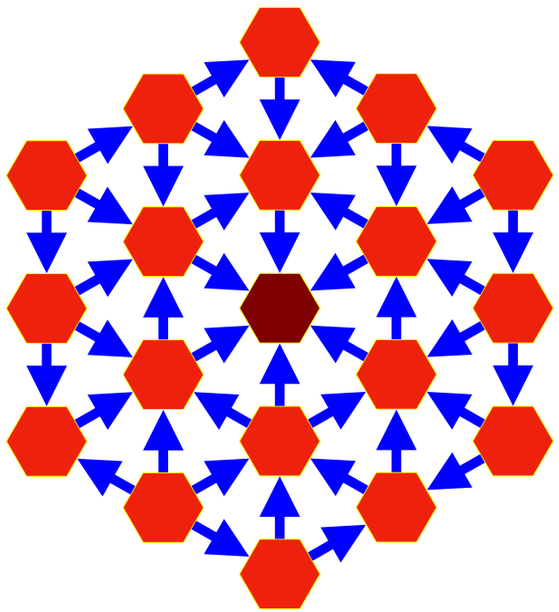


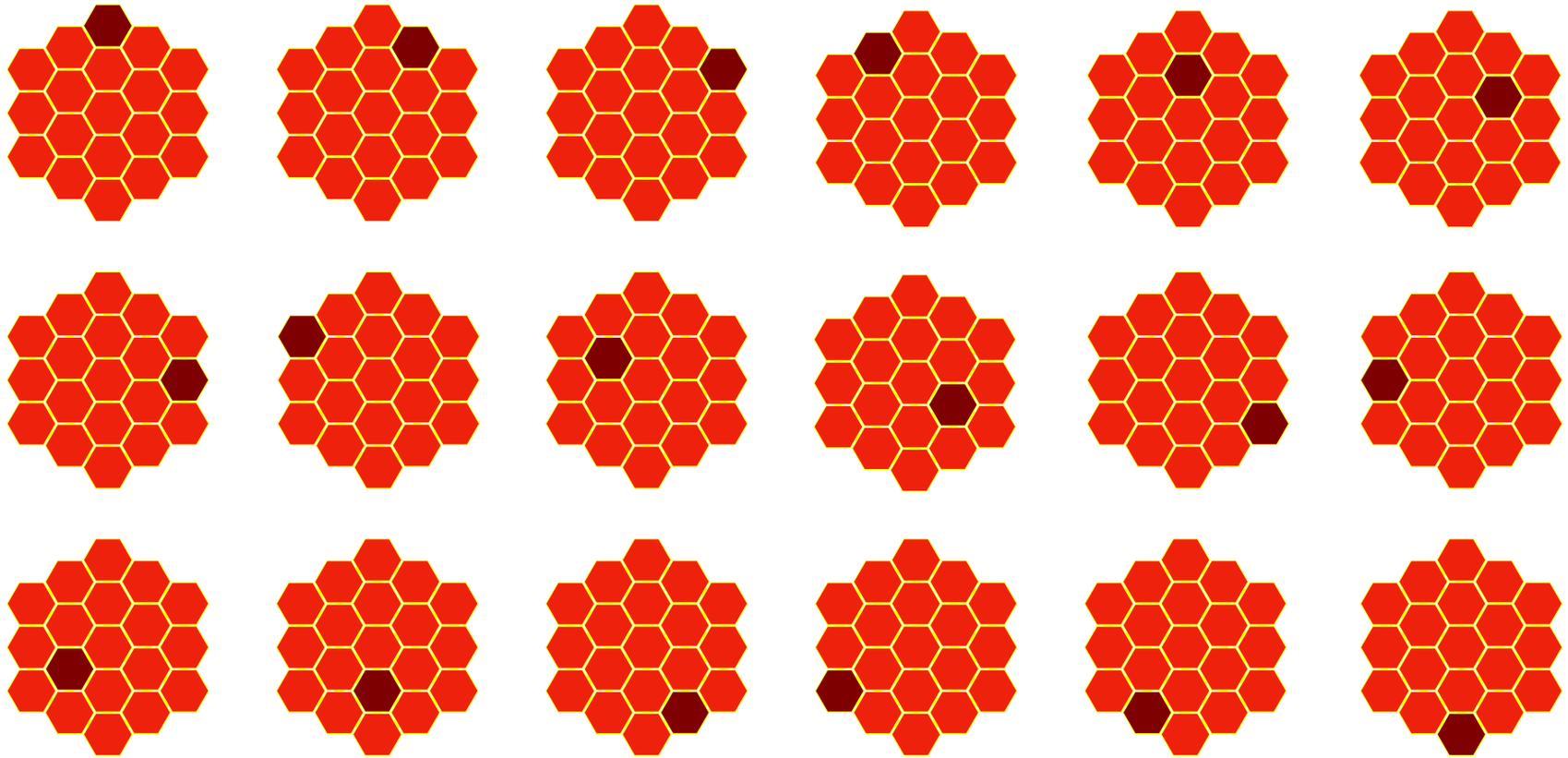
Fallible Chiplet Mesh





Spacecraft Arrays: In Formation*





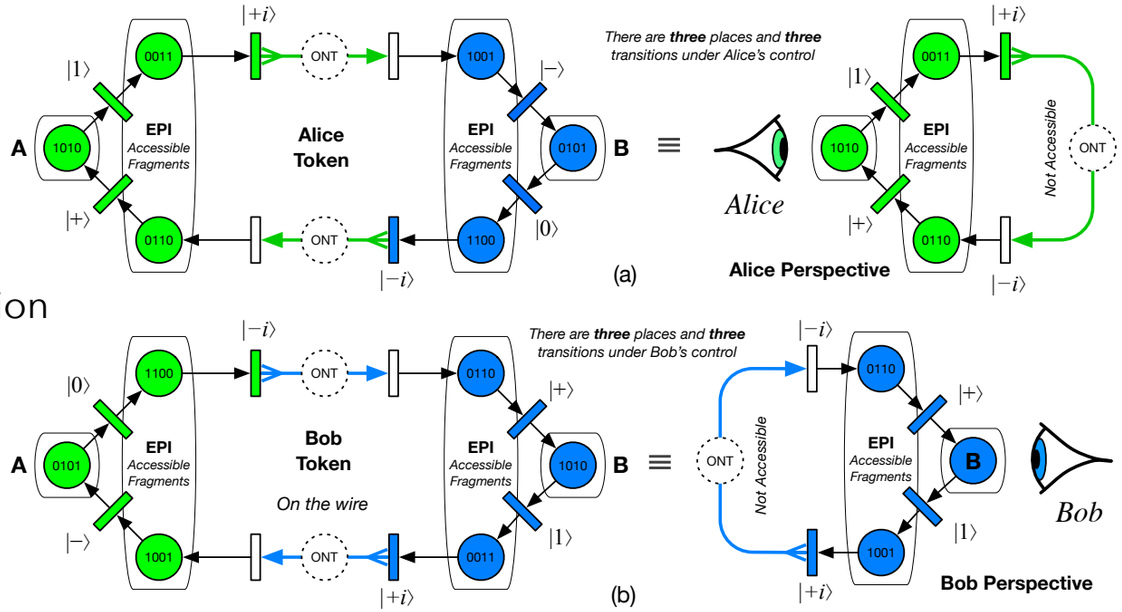
Complete Redundancy: Any cell can become a controller if others fail

*From: ItsAboutTime.club talk: Swarming Proxima Centauri: How Really Good Clocks Enable Optical Communication Over Interstellar Distances

Quantum Ethernet (Dual SAW-Petri-Spekkens-Protocol)

A software infrastructure for datacenter networks based on algorithms whose assumptions about causality go beyond Newtonian and Minkowski spacetime. We design and verify protocols for direct (near neighbor connected) networks that can be deployed on FPGA-enabled SmartNICs to address fundamental problems in distributed systems. This leads to a system of rewriting rules that can execute in multiway application fragments 'invisibly' and 'indivisibly' in the FPGA substructure cluster

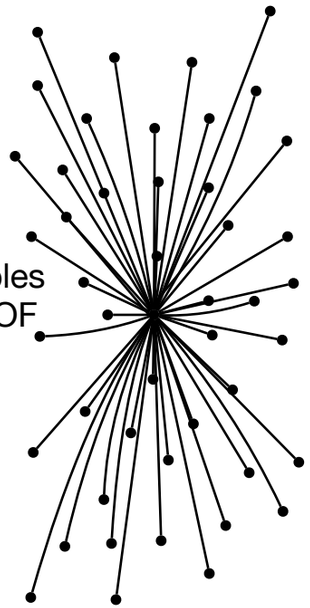
Compile Rust to Petri nets
For formal verification and simulation



Compile Petri nets to Verilog
For Deployment on FPGA's

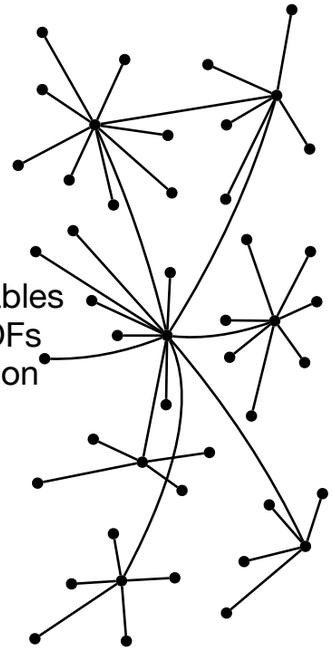
Centralized

Long cables
ONE SPOF



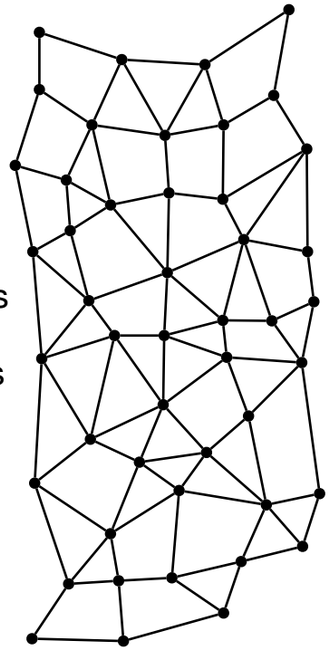
Decentralized

Long/Medium cables
MULTIPLE SPOFs
(Network partition possibilities)



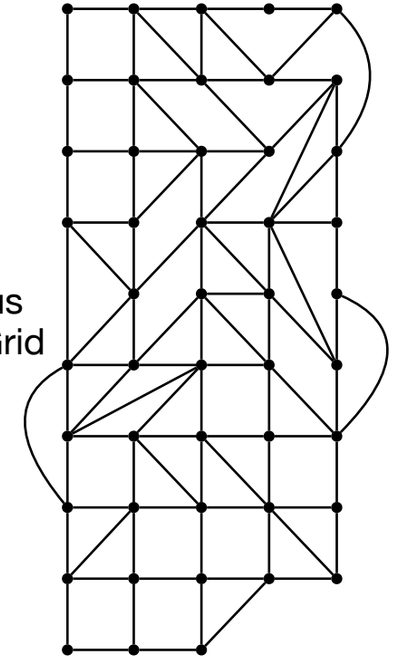
Distributed

Short local links
No SPOFs
No Bottlenecks

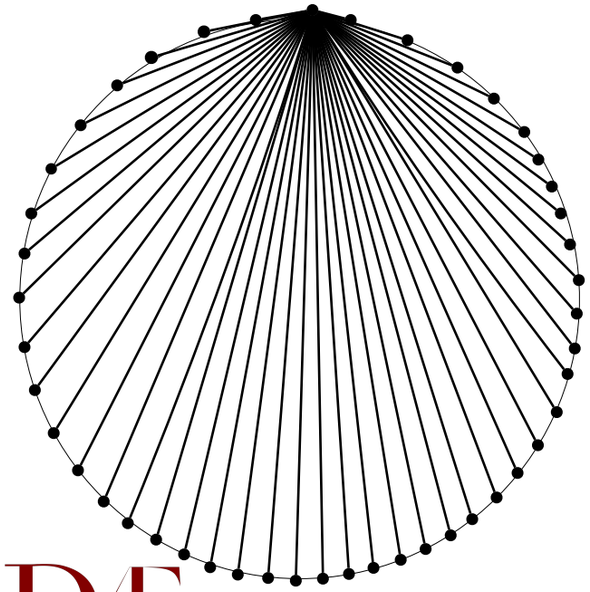


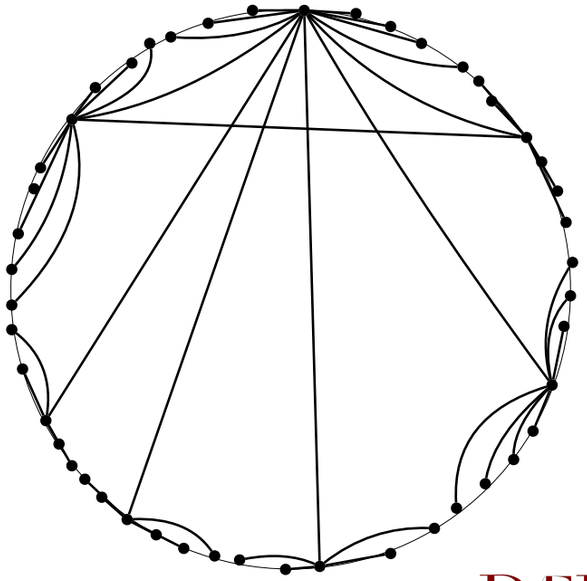
Daedaelus

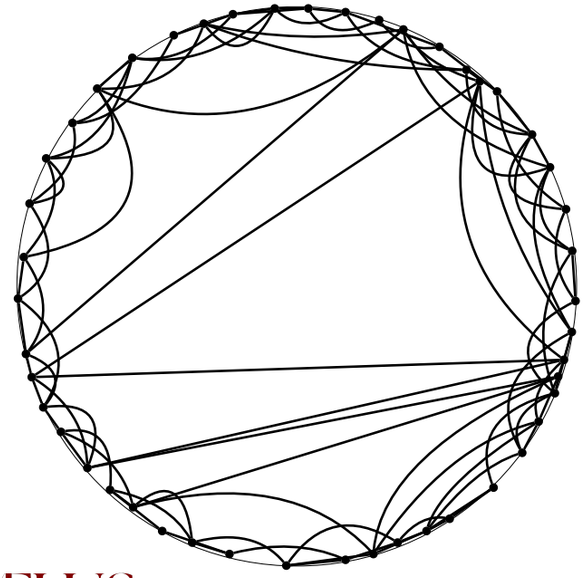
Daedaelus
Evolving Grid

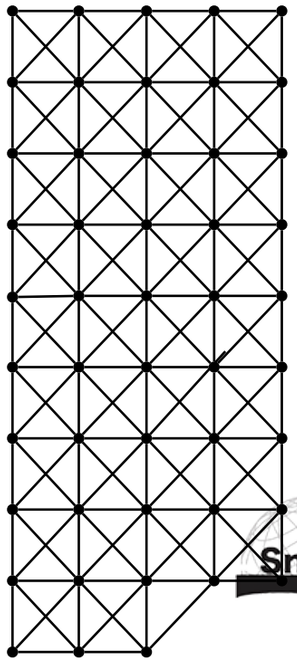


Valency 5









Valency 8

DÆ

DÆDÆLUS

SmartNICs SUMMIT

ItsAboutTime.club

- A relationship with time is intrinsic to everything we do within and between our networked computers.
- An assumption that time is a smooth, irreversible, global Newtonian/Minkowskian background is a common but rarely questioned belief in computer science; yet, physicists now know this model to be incorrect.
- Our guest speakers are all people who have thought deeply about the nature of time. We collectively realize that a new understanding could potentially revolutionize the way we approach physics, computer science, chemistry, neuroscience, and many other subjects.
- SmartNICs in Particular can benefit.
    - Temporal Intimacy with bits on the wire. Decoupled transactions, CAP: Consistency, Availability, Partitioning.

Home Episodes About Blog Contact

With Paul Borrill

It's About Time!

A place to discuss our evolving knowledge of the nature of time and causality. For physicists, computer scientists, mathematicians, neuroscientists, philosophers, and practicing engineers.

Listen To All Episodes →

Follow Paul Borrill on Twitter

Follow IAT! on Clubhouse

Follow Paul Borrill on Clubhouse

Spacetime is Doomed.

A place to discuss our evolving knowledge of the nature of time and causality.
For physicists, computer scientists, neuroscientists, philosophers and practicing engineers.

A relationship with time is intrinsic to everything we do within and between our networked computers. An assumption that time is a smooth, irreversible, global Newtonian/Minkowskian background is a common but rarely questioned belief in computer science; yet, physicists now know this model to be incorrect. Our guest speakers are all people who have thought deeply about the nature of time. We collectively realize that a new understanding could potentially revolutionize the way we approach physics, computer science, chemistry, neuroscience, and many other subjects.

SmartNIC Skyscraper Model

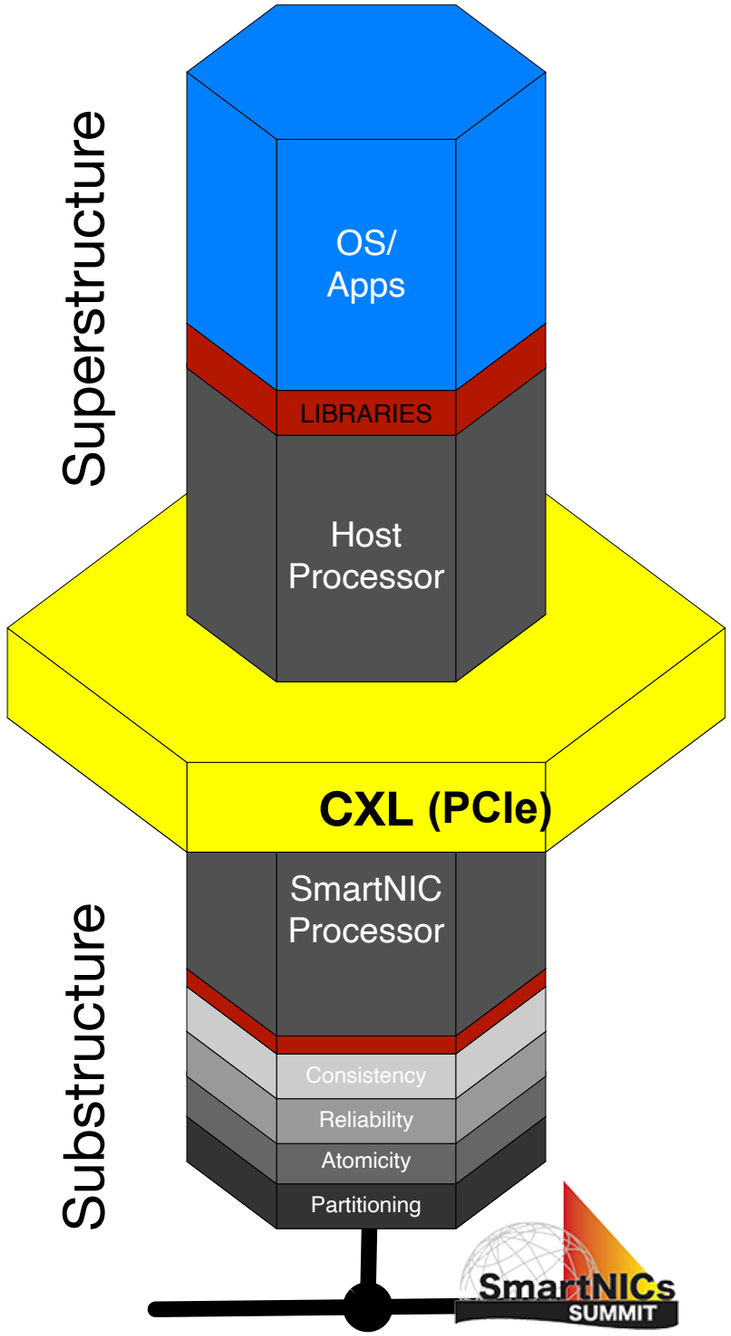
- New Substructure Consortium – like SNIA
- New IEEE Standard – for Distributed Systems
- Revolutionary Technology from DAEDAELUS:
  - Willing to Share, OpenSource*, License Fairly & Reasonably

https://github.com/JonathanGorard/Labyrinth

The Revolution Starts here: at the SmartNICs Summit

Meet on Wednesday Afternoon

Contact: info@daedaelus.com



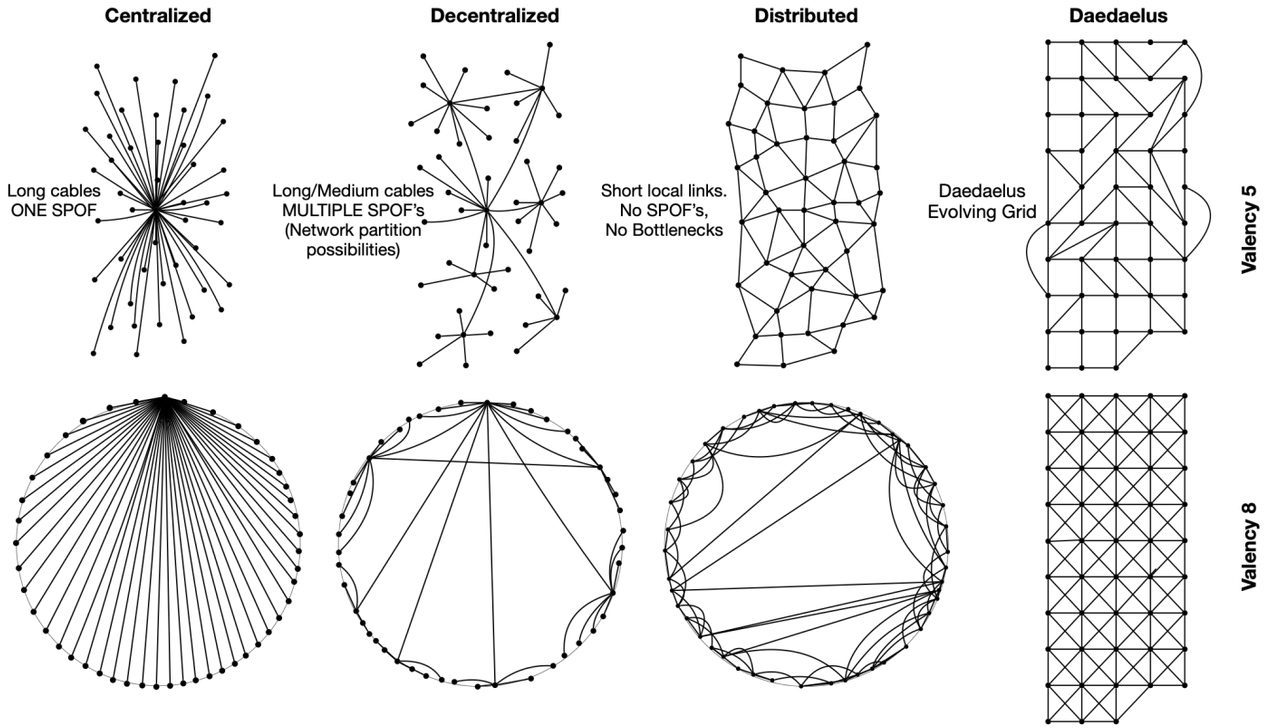
DÆDÆLUS, Inc

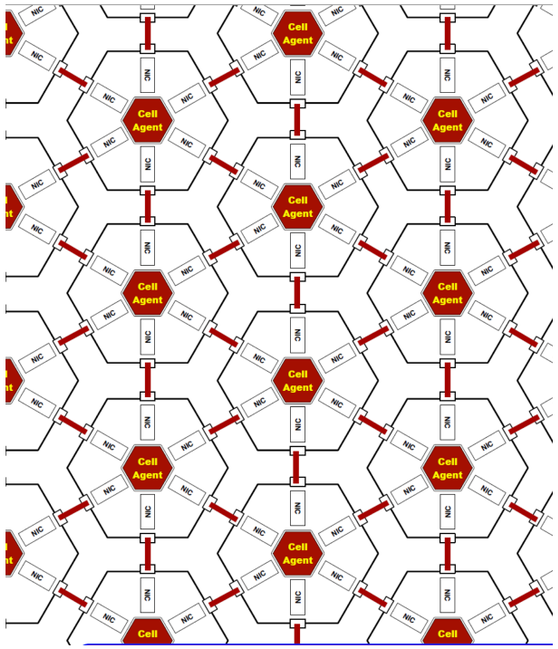
Labyrinh for SmartNICs

We solve fundamental problems in distributed systems using protocols, data structures and algorithms inspired by Quantum Information Theory and Multiway Systems. Our market is next generation platforms for secure, reliable, distributed computing on the edge. We provide Microdatacenters with a fundamentally more reliable and programmable graph infrastructure. Initial use-cases include Transaction systems, Digital Twins, AI/ML/LLM infrastructure, Multiplayer Games and Interface to Quantum Computers.

DÆ

Distributed Autonomous Ethernet





Distributed Systems API's for SmartNICs

- GREY FAILURE NEEDS OUR HELP
- KUBERNETES NEEDS OUR HELP
- SECURE ENCLAVES NEED OUR HELP

- THE CAP THEOREM NEEDS OUR HELP
- PROGRAM WITH LINGUA FRANCA
- TRANSACTIONS NEED OUR HELP

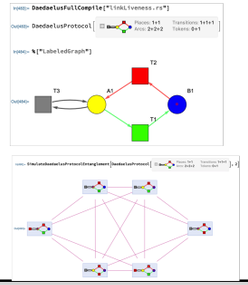
IN HONOR OF MARK CARLSON

POC Table Top Demo in the Exhibit Hall Session: Wednesday, 04:20-5:20 PM x

Labyrinth - Formal Verification & Simulation Pipeline

1. Compile any component of the Daedaelus protocol suite directly from Rust into our symbolic protocol description language, based on a generalization of colored Petri nets
2. Components can be simulated fully (with interactions with other components), allowing us to compute the state transition graphs exhibiting all possible (non-deterministic) evolution histories for the overall protocol
3. We extract *entanglement* information from these protocols, illustrating which *microstates* of the protocol are *entangled* (related by non-Cartesian tensor product, and hence non-separable), as in quantum mechanics.
4. We can also simulate typical failure scenarios, such as link failure or packet loss, and quantify the protocol's robustness and ability to recover

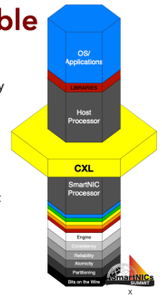
https://github.com/JonathanGorard/Labyrinth



We make Transactions Reliable

- Reliable Substructure Clusters
  - Reversible transfers at line rates, with ultra-low latency
- No tradeoff of Bandwidth and Latency
- No Metastable Failures

- **FPGA's do not have a halt state!**
  They just run. Just circuits. Stuff goes in comes out, no halt states in between
- Unlike ASICS, they don't take years to create



DÆ DÆDÆLUS x

DÆ

SmartNICs SUMMIT