ASIC Clouds: Specializing the Datacenter

Ikuo Magaki+, Moein Khazraee, Luis Vega Gutierrez, and Michael Bedford Taylor

UC San Diego

+UC San Diego, Toshiba

Presented at ISCA 2016.

Compute Trends in 2016

Bifurcation of computation into Client and Cloud

- Client is mobile SoC
- Cloud is implemented by datacenters

Compute Trends in 2016

Bifurcation of computation into Client and Cloud

- Client is mobile SoC
- Cloud is implemented by datacenters

End of Dennard Scaling

- Rise of Dark Silicon[1]
- Dark Silicon-aware design techniques[2]
    - Specialization (accelerators)
    - Low voltage or Near-threshold operation

[1] "Conservation Cores", ASPLOS 2010; GreenDroid, HOTCHIPS 2010.
[2] "A Landscape of the Dark Silicon Design Regime", Taylor, IEEE Micro 2013.

Compute Trends in 2016

Bifurcation of computation into Client and Cloud

- Client is mobile SoC
- Cloud is implemented by datacenters

End of Dennard Scaling

- Rise of Dark Silicon
- Dark Silicon-aware design techniques
  - Specialization (accelerators)
  - Low voltage or Near-threshold operation

Early Signs of Specialization in the Datacenter

GPU-based clouds

- Deep Neural Networks [Baidu Minwa]

Early Signs of Specialization in the Datacenter

GPU-based clouds

- Deep Neural Networks [Baidu Minwa]

FPGA-based clouds

- Hedgefund Portfolio Evaluation [JP Morgan]
- High Frequency Trading [Most Wall Street firms]
- Catapult [Microsoft]

Early Signs of Specialization in the Datacenter

- GPU-based clouds
  - Deep Neural Networks [Baidu Minwa]
- FPGA-based clouds
  - Hedgefund Portfolio Evaluation [JP Morgan]
  - High Frequency Trading [Most Wall Street firms]
  - Catapult [Microsoft]
- Xeon Processors
  - Customer specialized SKUs [Oracle]
  - Xeon-D [Facebook]

Early Signs of Specialization in the Datacenter

GPU-based clouds

- Deep Neural Networks [Baidu Minwa]

FPGA-based clouds

- Hedgefund Portfolio Evaluation [JP Morgan]
- High Frequency Trading [Most Wall Street firms]
- Catapult [Microsoft]

Xeon Processors

- Customer specialized SKUs [Oracle]
- Xeon-D [Facebook]

What about ASIC-based clouds?

ASIC Clouds: Key Motivation

The Cloud model leads to growing classes of planet-scale computations which incur high Total Cost of Ownership costs for the provider

e.g. FB runs face rec on 2B pics/day
Siri recognizes speech for ~1 Billion iOS users
YouTube performs Video Transcoding for uploads (to Google VP9)

ASIC Clouds: Key Motivation

The Cloud model leads to growing classes of planet-scale computations which incur high Total Cost of Ownership for the provider

e.g. FB runs face rec on 2B pics/day
Siri recognizes speech for ~1 Billion iOS users
YouTube performs Video Transcoding for uploads (to Google VP9)

These computations are *scale-out*, not *scale-up* computations, so we are doing the same computation across millions or billions of users.

ASIC Clouds: Key Motivation

The Cloud model leads to growing classes of planet-scale computations which incur high Total Cost of Ownership for the provider

e.g. FB runs face rec on 2B pics/day
Siri recognizes speech for ~1 Billion iOS users
YouTube performs Video Transcoding for uploads (to Google VP9)

These computations are *scale-out*, not *scale-up* computations, so we are doing the same computation across millions or billions of users.

As these computations become sufficiently large, we can specialize the hardware for that particular computation to reduce TCO.

ASIC Clouds: Efficiently Deploying Accelerators into Datacenters

ASIC Cloud: Purpose-built datacenter comprising large arrays of accelerators (like those proposed at ISCA) packed hierarchically into chips, PCBs, and then racks.

ASIC Clouds: Efficiently Deploying Accelerators into Datacenters

ASIC Cloud: Purpose-built datacenter comprising large arrays of accelerators (like those proposed at ISCA) packed hierarchically into chips, PCBs, and then racks.

The Paper's Results:

- *<u>Huge</u> benefits to specializing servers for the accelerator*
  - *Removing unneeded general-purposeness*
  - *We optimize Silicon, PCB, Thermals, Power Delivery, Cooling, Voltage*
- *<u>Significant</u> TCO benefits if the workload is large enough*
  - *Reduction in power-related costs*
  - *Reduction in marginal HW cost*

ASIC Clouds: Efficiently Deploying Accelerators into Datacenters

ASIC Cloud: Purpose-built datacenter comprising large arrays of accelerators (like those proposed at ISCA) packed hierarchically into chips, PCBs, and then racks.

The Paper's Results:

<u>Huge</u> benefits to specializing servers for the accelerator

Removing unneeded general-purposeness

We optimize Silicon, PCB, Thermals, Power Delivery, Cooling, Voltage

<u>Significant</u> TCO benefits if the workload is large enough

Reduction in power-related costs

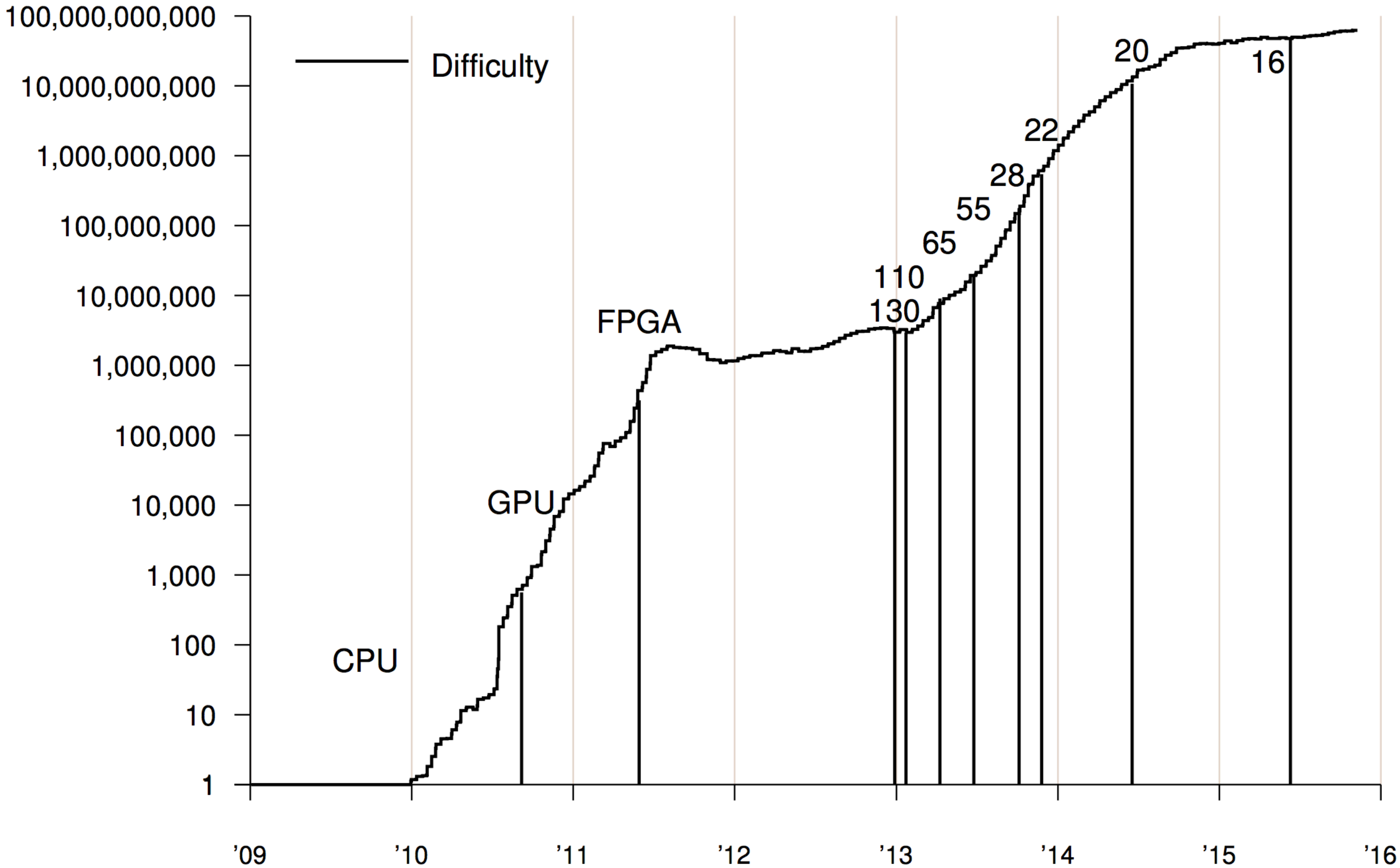
Reduction in marginal HW cost

Going ASIC Cloud will become
a <u>routine business decision</u> because it saves money!

ASIC Clouds Exist Today

I'm not making this up…

16 NM



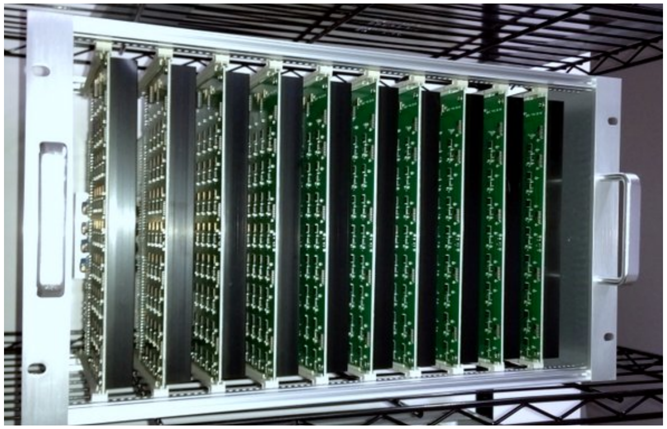
World Bitcoin Mining Capacity and ASIC Node

ASIC Clouds Exist Today

ASIC Clouds for Bitcoin mining have hit 300-500 MW worldwide.

Current throughput is > <u>1.2 Billion GPUs</u> (!)
(Some machines are equivalent to 8500 GPUs)

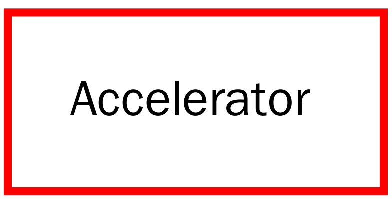
For this paper, I purchased 8 different bitcoin miners, and reverse engineered them.
Many were deeply suboptimal.



Come by San Diego to see my museum!

ASIC Cloud Architecture

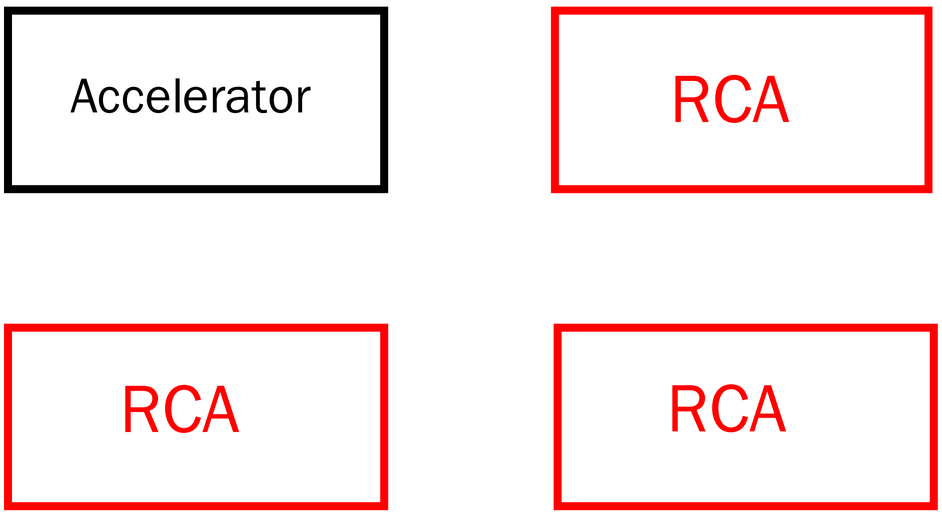
We propose a prototypical architecture for all ASIC clouds....



It all starts with an accelerator for a planet-scale computation.
Maybe it's a commercial IP core, or custom designed widget in Verilog.

ASIC Cloud Architecture

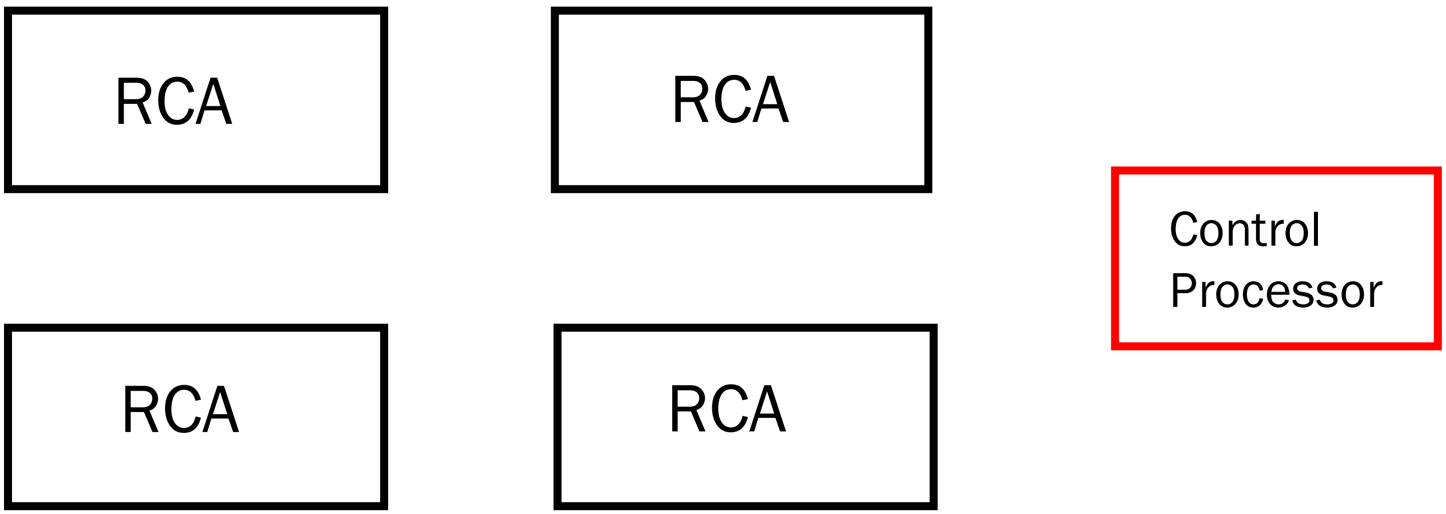
We propose a prototypical architecture for all ASIC clouds....



Replicate this accelerator multiple times inside an ASIC die.
We'll now call it a "replicate compute accelerator", or "RCA".

ASIC Cloud Architecture

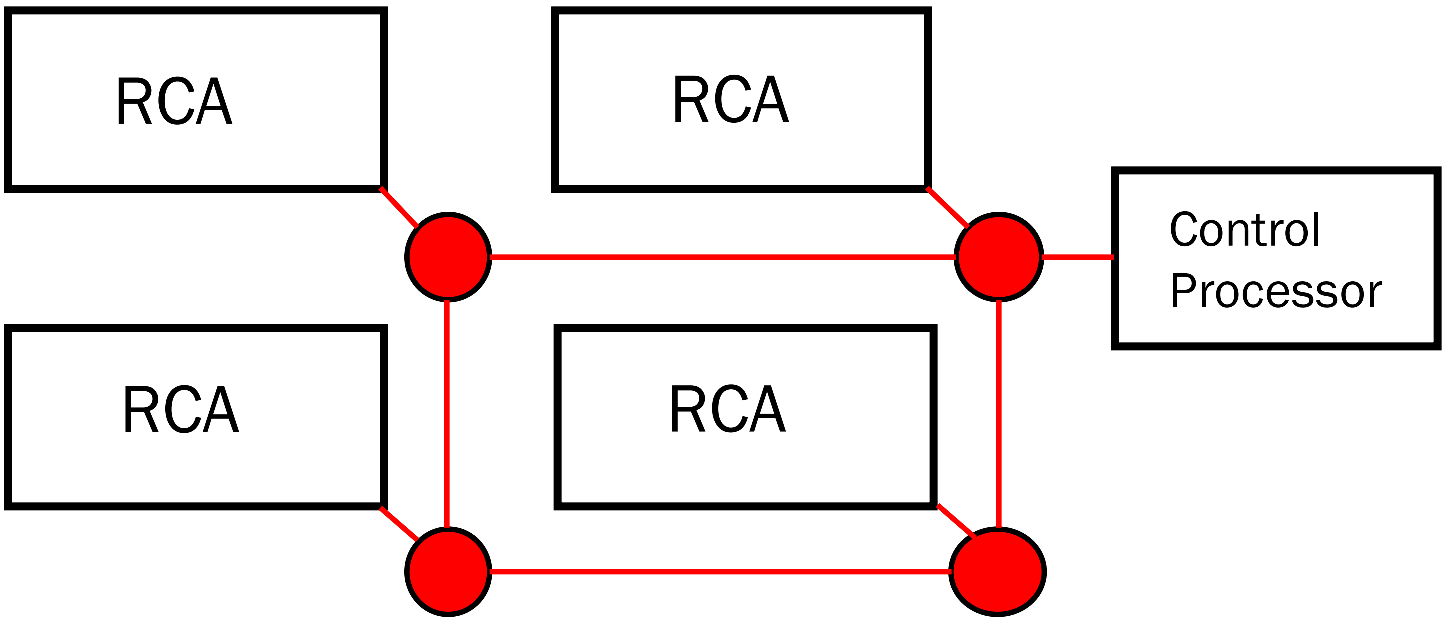
We propose a prototypical architecture for all ASIC clouds....



Then we add a control processor to distribute work and schedule computation onto the RCAs.

ASIC Cloud Architecture

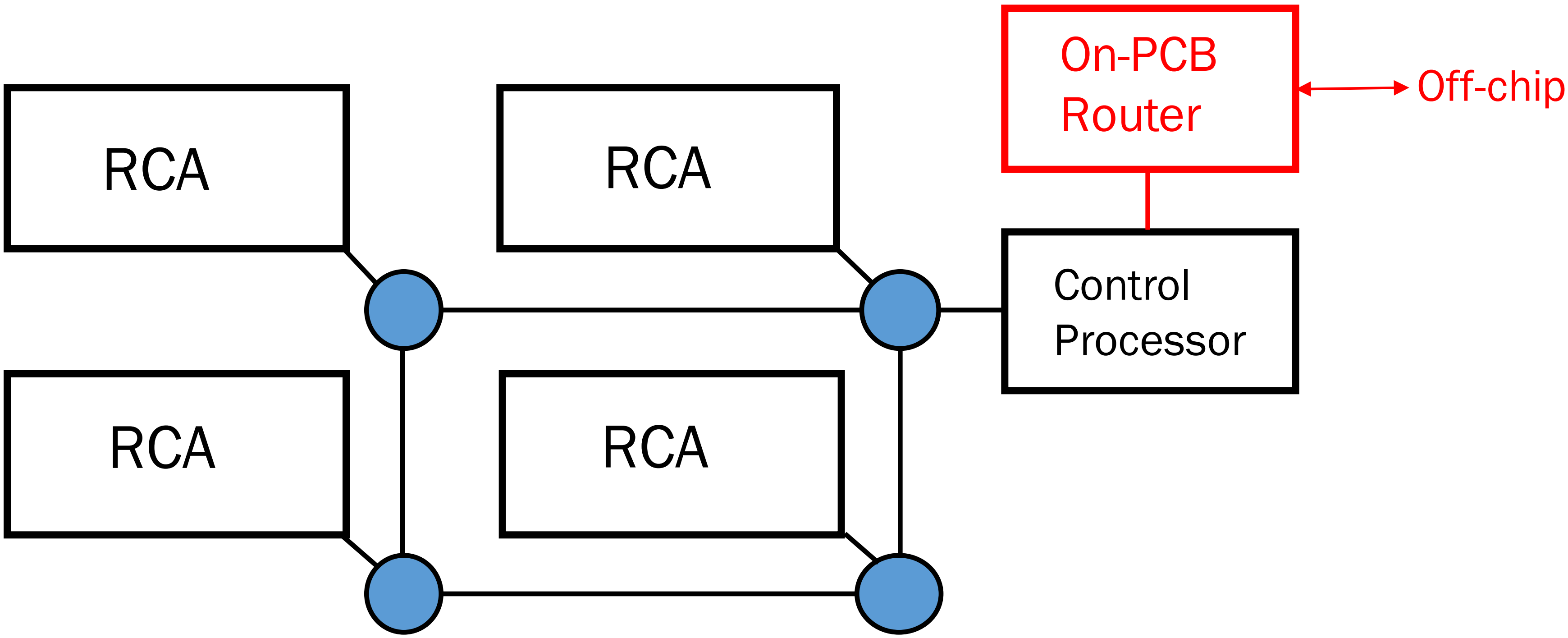
We propose a prototypical architecture for all ASIC clouds….



*Work is distributed over a very simple on-chip network, the **On-ASIC Network**, which is provisioned according to the needs of the RCAs.*
RCA's usually do not talk to each other.

ASIC Cloud Architecture

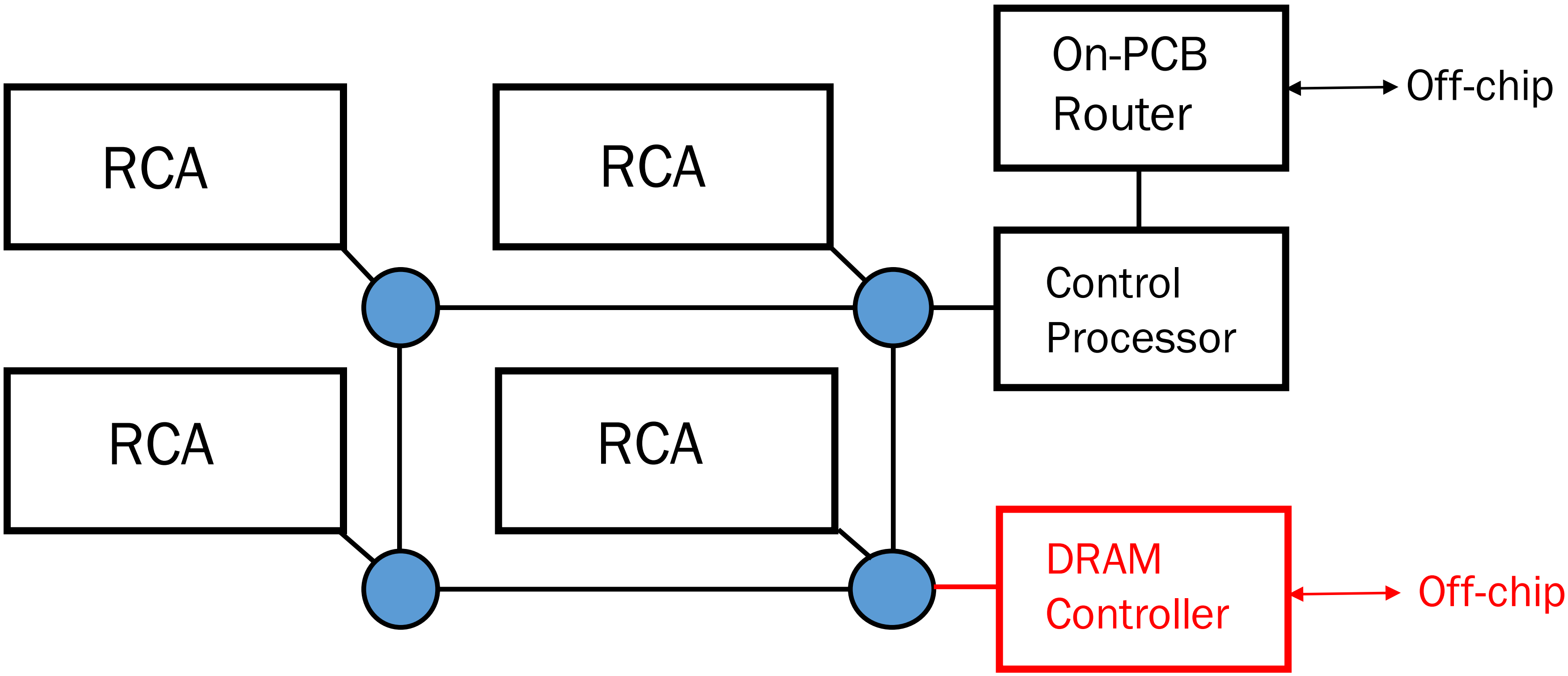
We propose a prototypical architecture for all ASIC clouds....



The control processor receives work from off-chip via the On-PCB router.

ASIC Cloud Architecture

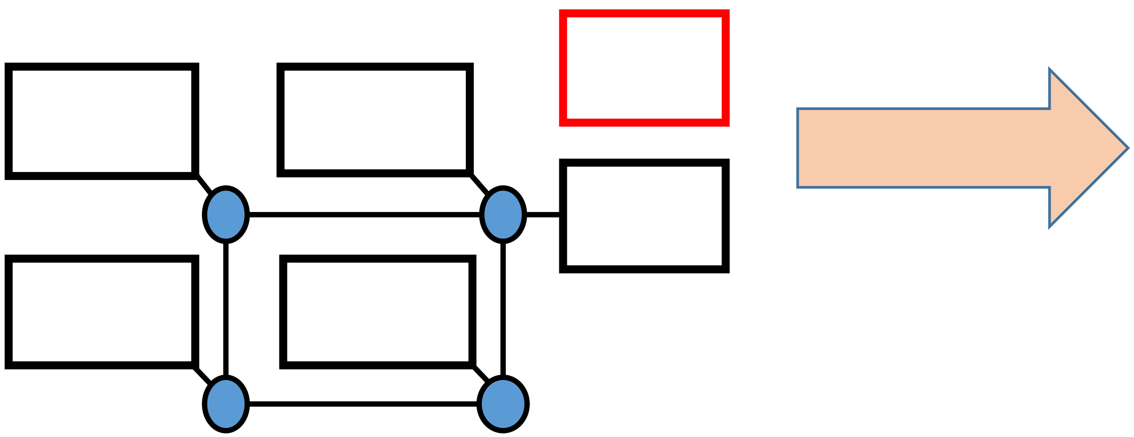
We propose a prototypical architecture for all ASIC clouds….

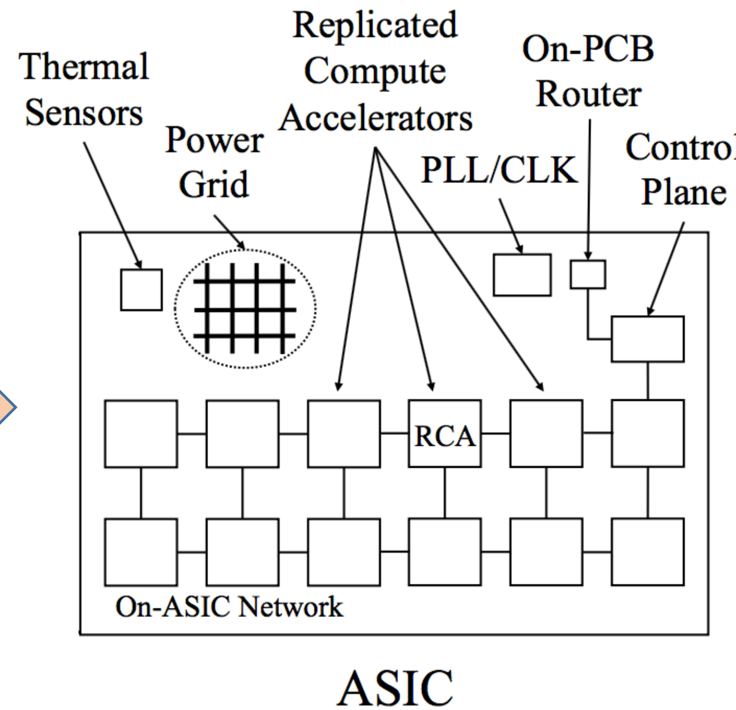


For those accelerators that need off-chip DRAM, we add shared DRAM controllers

ASIC Cloud Architecture

We propose a prototypical architecture for all ASIC clouds....

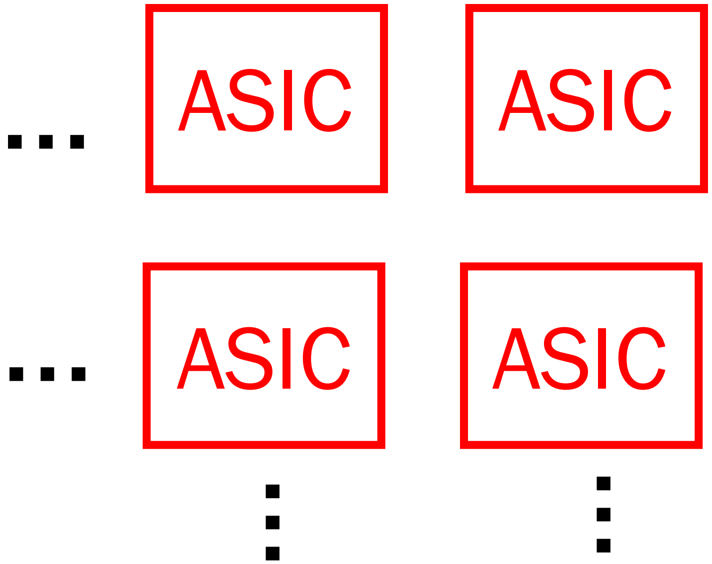




Bake it into an ASIC: PLL, Clock Tree, Power Grid, Flip Chip BGA Packaging...

ASIC Cloud Architecture

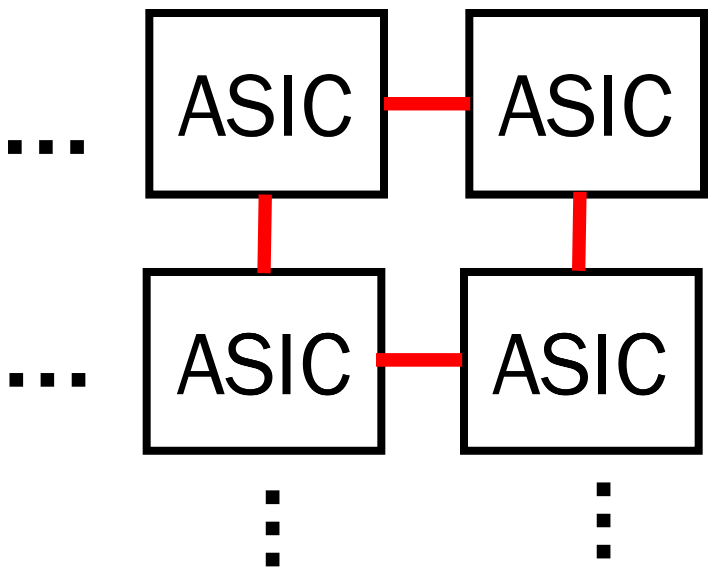
We propose a prototypical architecture for all ASIC clouds....



Then build the PCB by replicating ASICs across the PCB

ASIC Cloud Architecture

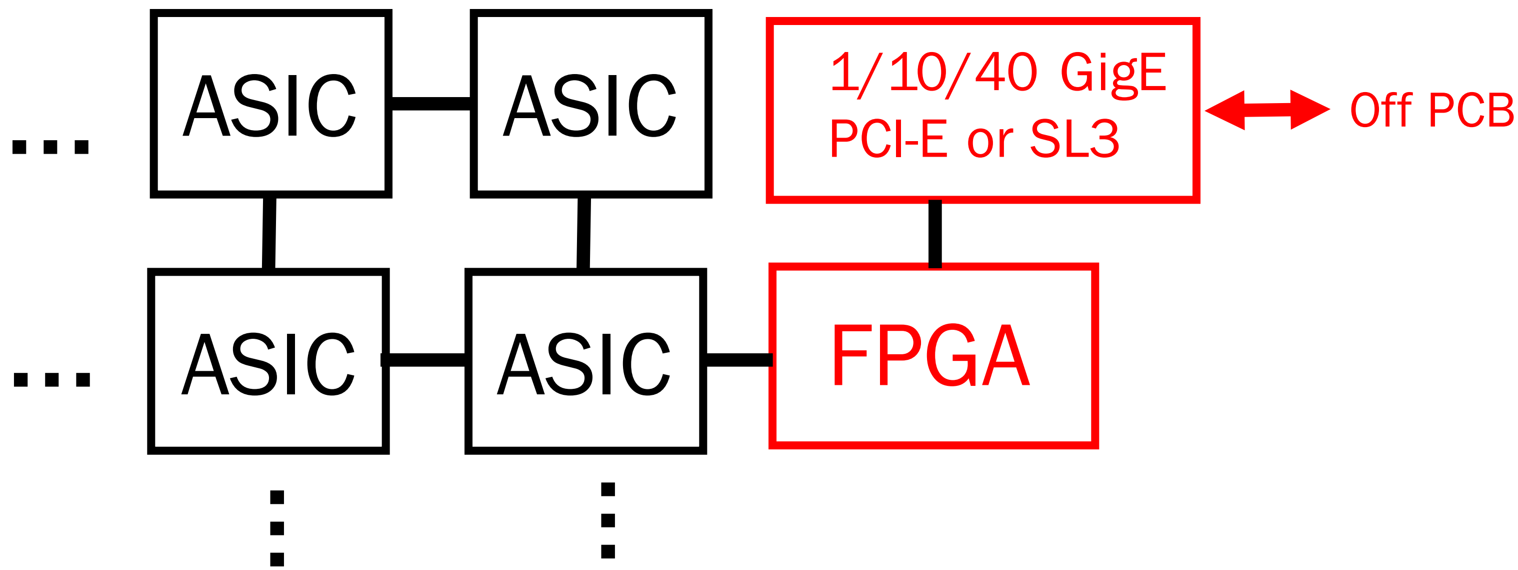
We propose a prototypical architecture for all ASIC clouds….



Connect their on-PCB routers via PCB traces

ASIC Cloud Architecture

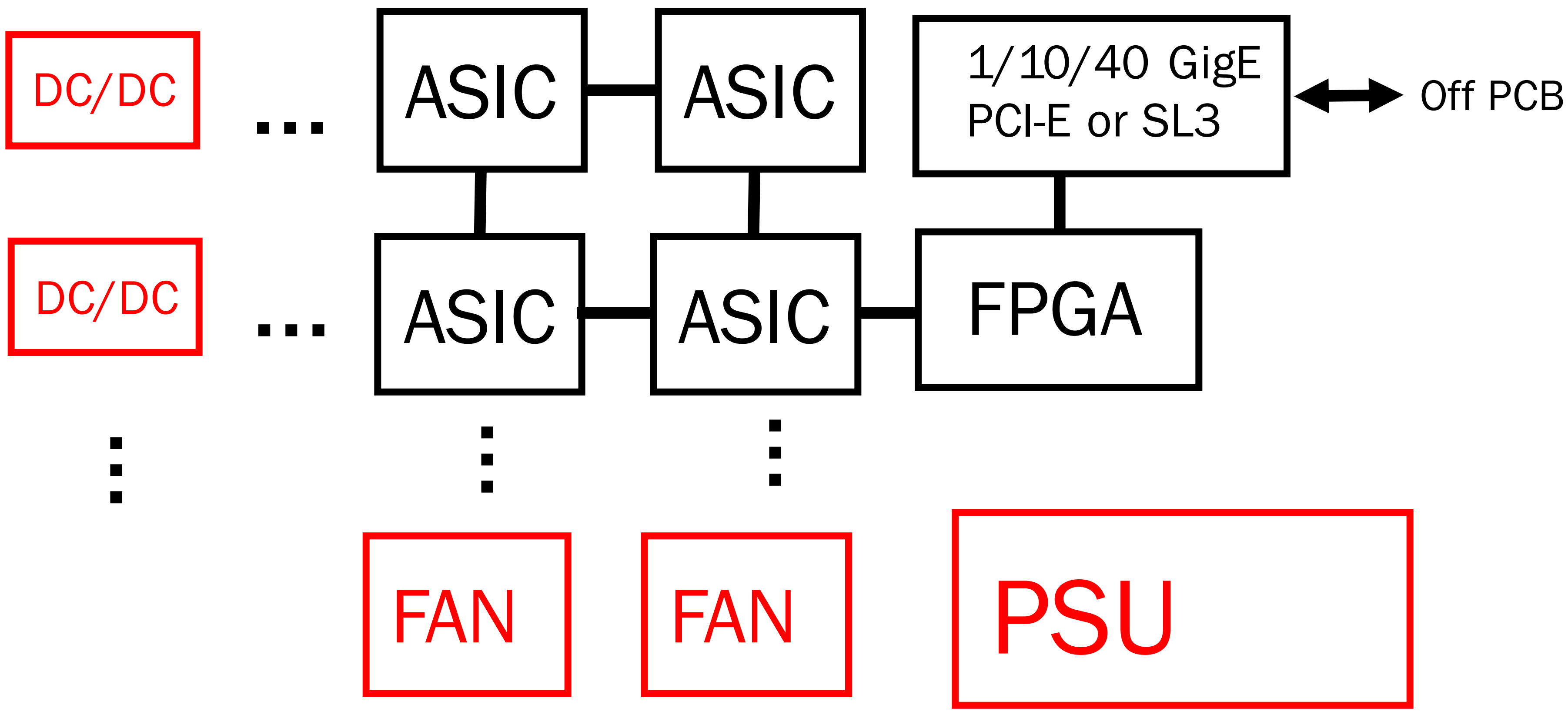
We propose a prototypical architecture for all ASIC clouds....



Connect the on-PCB network to an FPGA that routes data from off-PCB interface (e.g. GigE, PCI-E or SL3)

ASIC Cloud Architecture

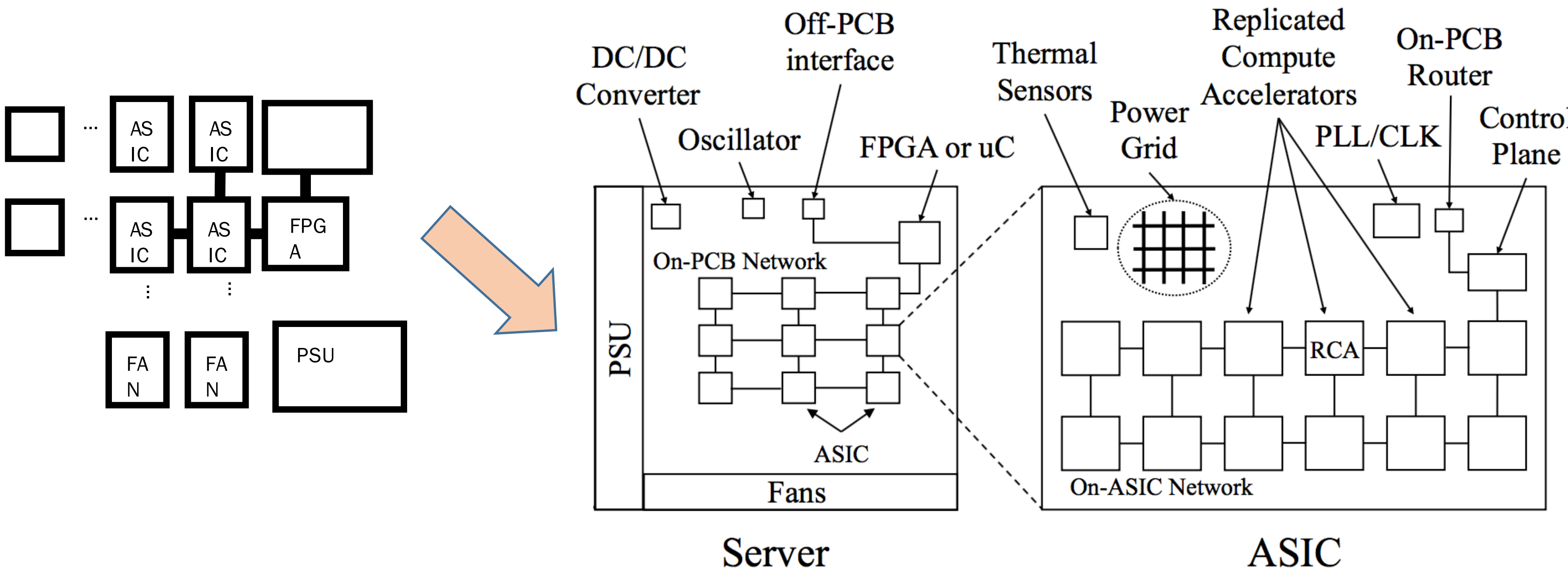
We propose a prototypical architecture for all ASIC clouds....



Then we add the plumbing: DC/DC, Fans, Heatsinks and PSU.

ASIC Cloud Architecture

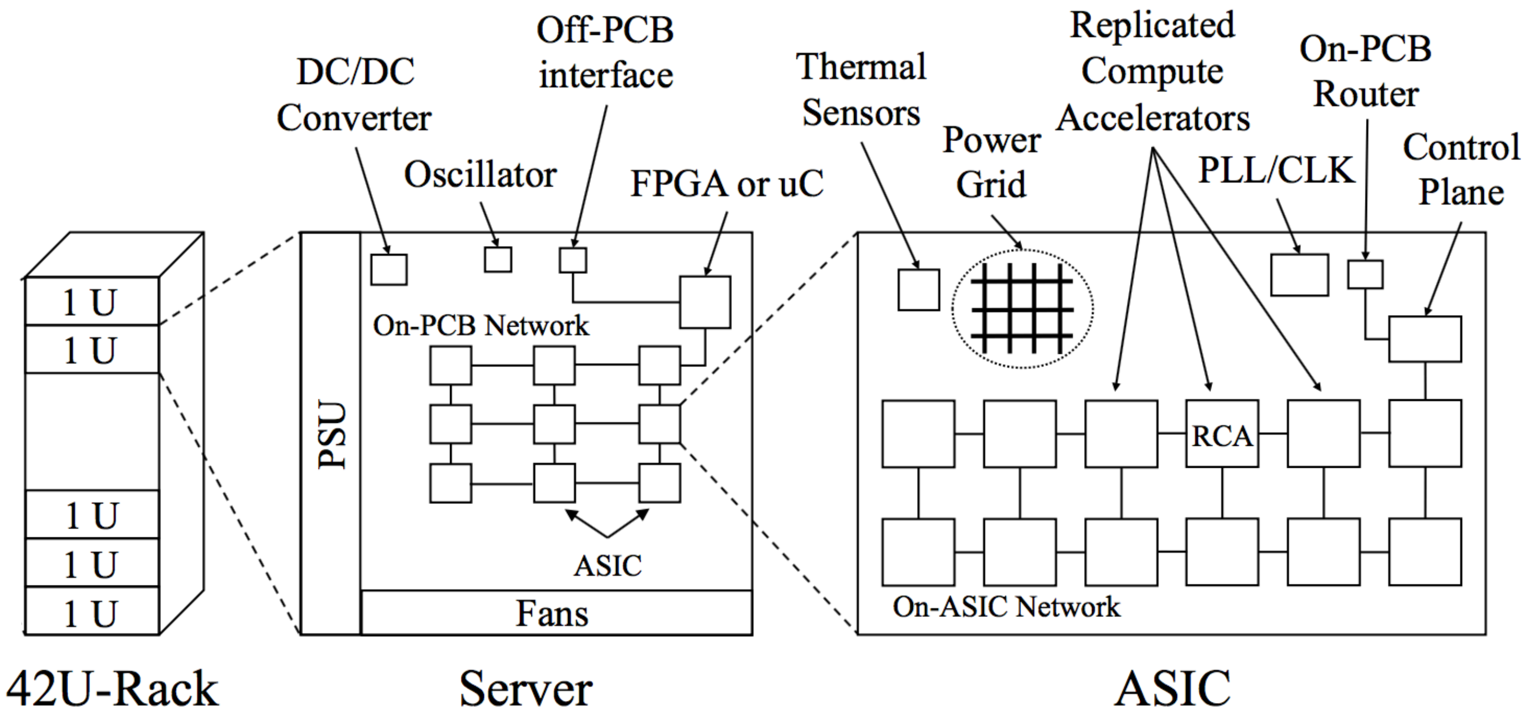
We propose a prototypical architecture for all ASIC clouds....



The PCB goes inside the chassis and we have an ASIC cloud server.

ASIC Cloud Architecture

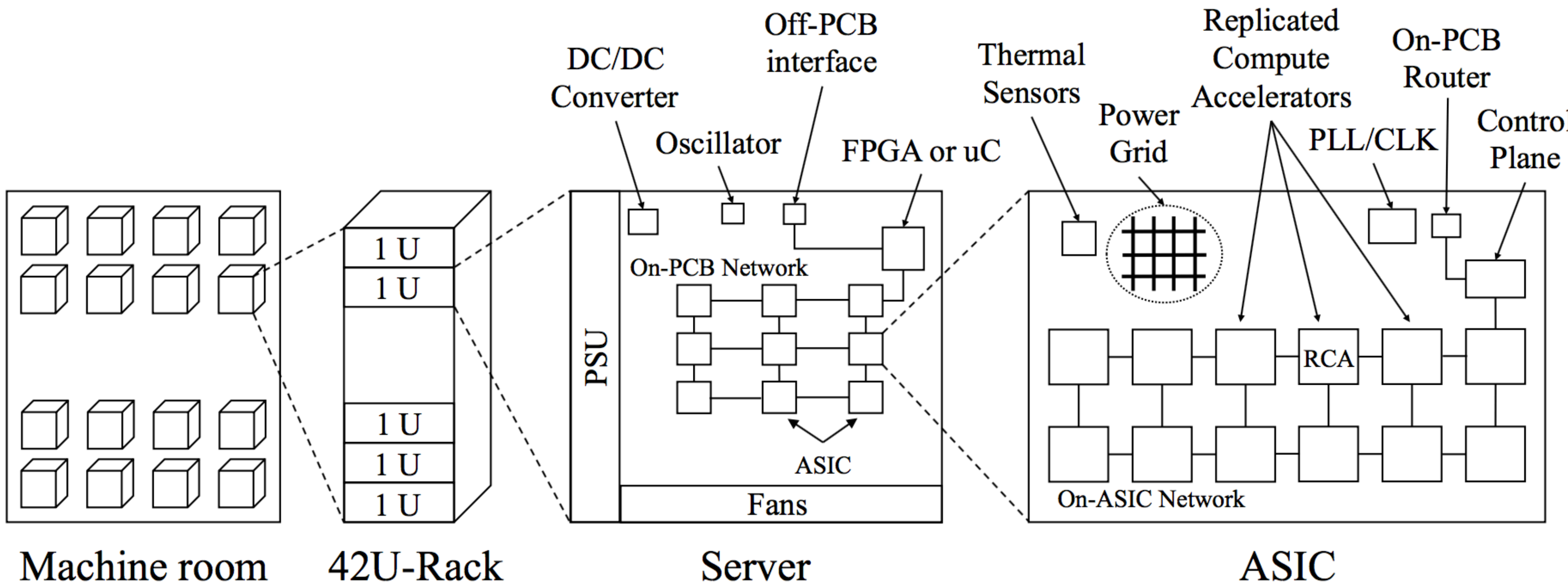
We propose a prototypical architecture for all ASIC clouds....



Servers are packed into standard 42U racks.

ASIC Cloud Architecture

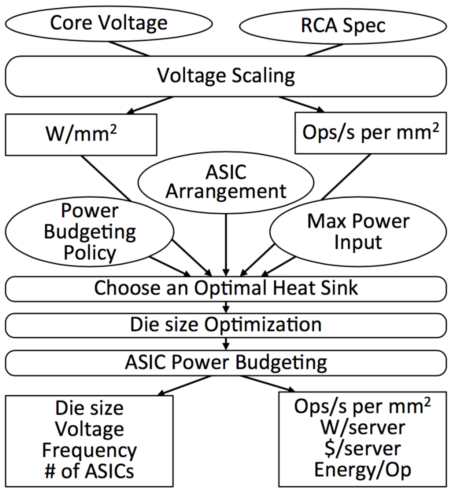
We propose a prototypical architecture for all ASIC clouds....

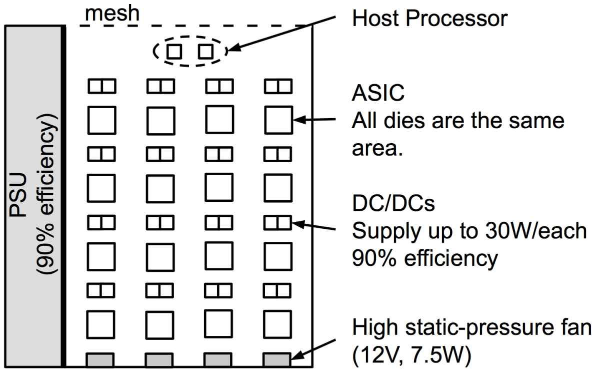


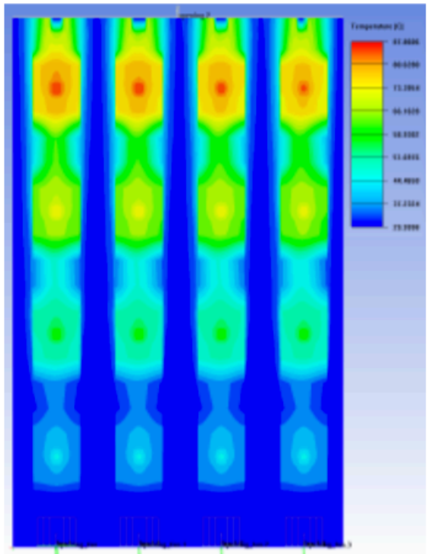
Racks are integrated into machine room.
In this paper, we do not specialize the machine room

(There's an interesting reason, see the paper.)

Complete Design Methodology from Verilog to TCO-Optimized Datacenter







Voltage selection,

Power supply design

Thermal Optimization

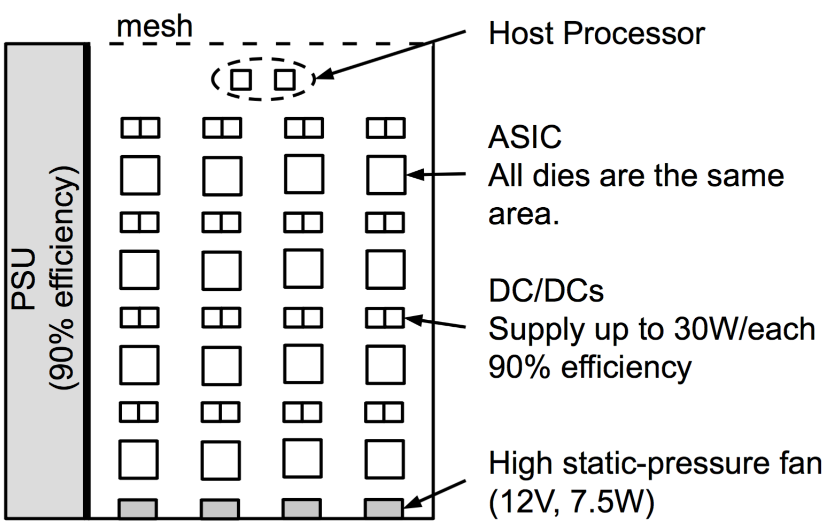
Complete Thermal Analysis using CFD

Papers shows how to take a ball of Verilog for an accelerator and turn it into a TCO-optimal ASIC Cloud…

(For time constraints, we highlight just a few items in the talk.. See the paper!)

ASIC Server Thermal Optimization

Using Computational Fluid Dynamics simulation

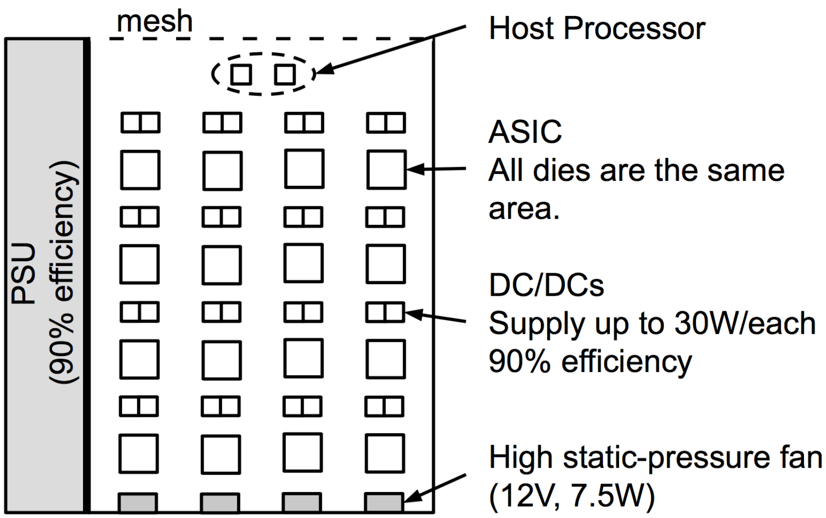


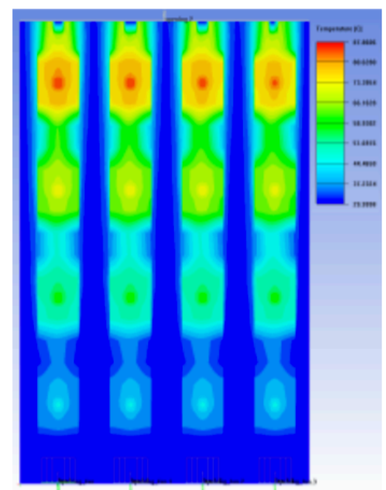
Physical Modeling with Ansys Icepak.

Each flip chip ASIC has a heatsink, which we optimize (# fins, width, materials and depth)
DC/DCs are on backside of PCB for space. Heatsink opt. depends on fan physics.

ASIC Server Thermal Optimization

Using Computational Fluid Dynamics simulation





Airflow

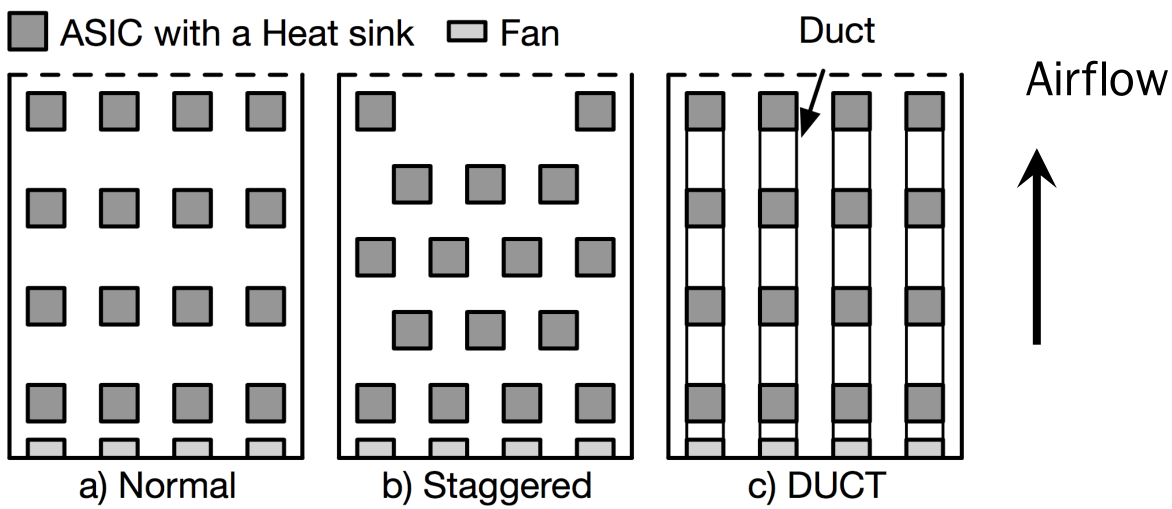
Physical Modeling with Ansys Icepak.

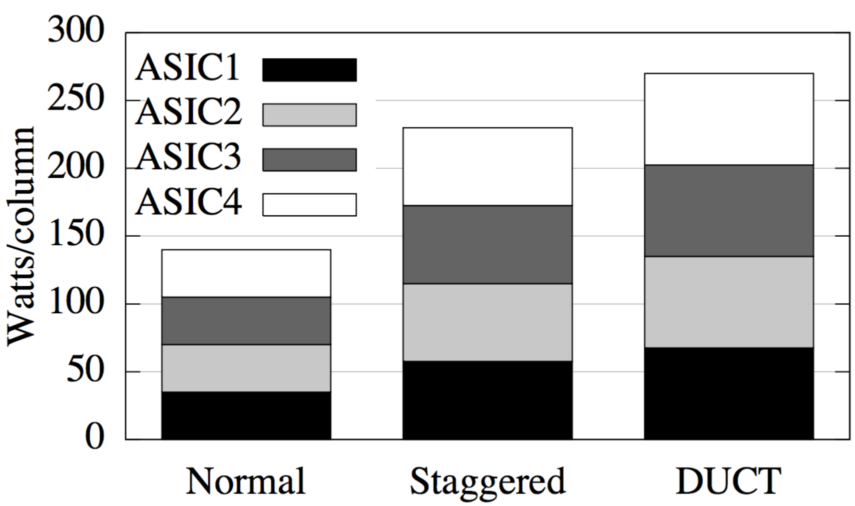
Each flip chip ASIC has a heatsink, which we optimize (# fins, width, materials and depth)
DC/DCs are on backside of PCB for space. Heatsink opt. depends on fan physics.

Rear ASIC is the thermally limiting one, because the hottest air blows over it.

ASIC Placement: Duct Wins

Server is optimized to maximize power under fixed max temp on ASICs.





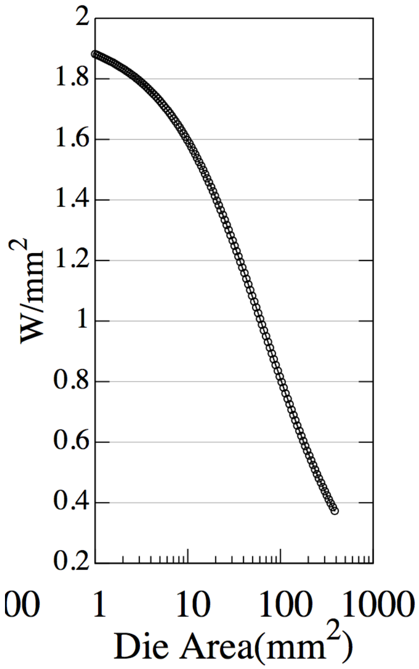
"Normal": hotspots are aligned; hottest air blows over hottest spots.

"Staggered" avoids this problem; much better

DUCT is even better than Staggered because less cold air "sneaks by."

How many RCAs per ASIC?

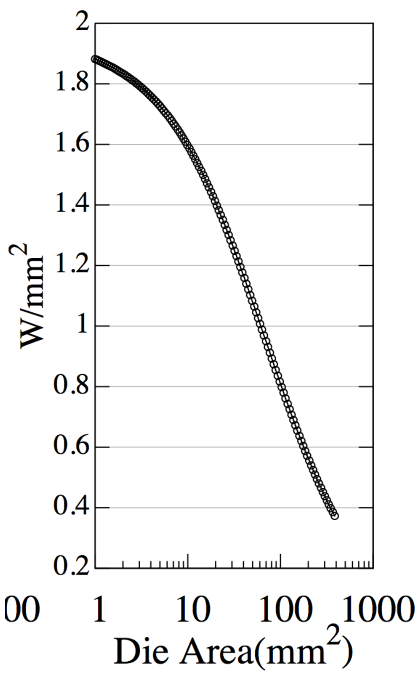
How does cooling ability change with die size?

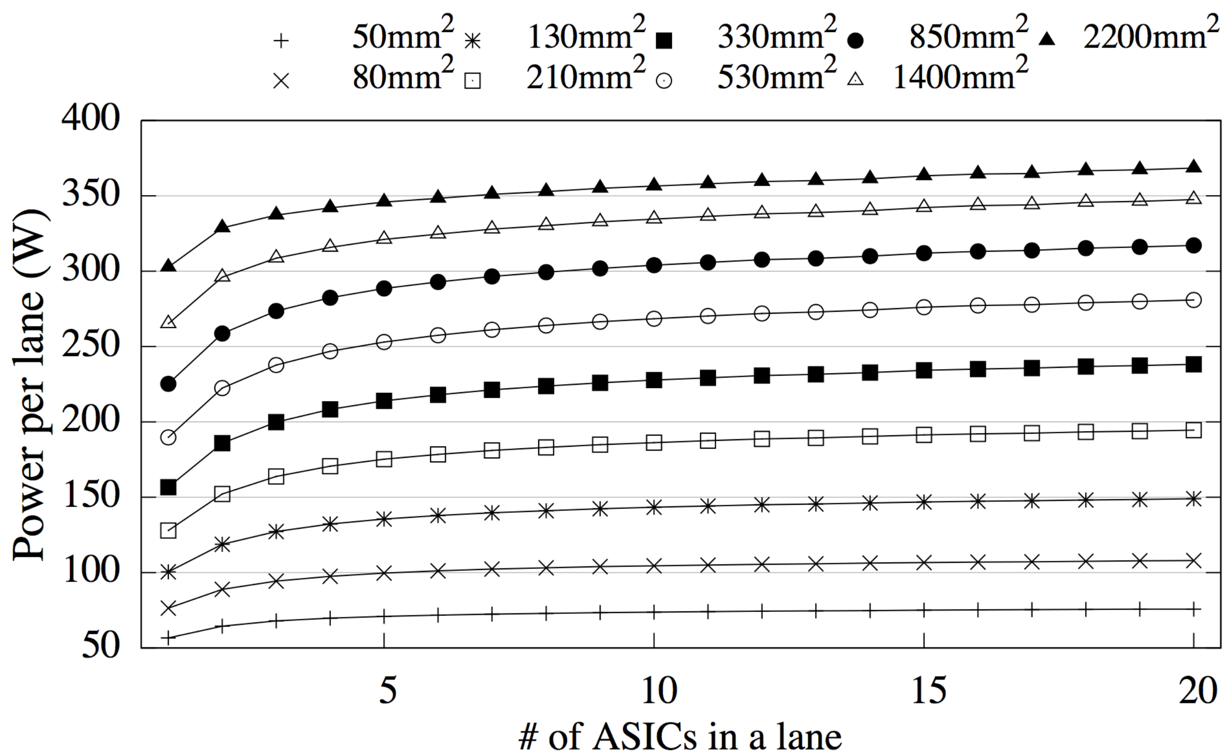


Smaller dies can sustain high power densities because heat crowds less and hotspots are cooler.

How many RCAs per ASIC?

How does cooling ability change with die size?





Smaller dies can sustain high power densities because heat crowds less and hotspots are cooler.

So given the same amount of silicon per lane, dividing it into more chips allows for more compute per lane.

ASIC Cloud Design: Key Metrics

How do we reason about optimality?

Typical Accelerator Metrics in ISCA papers:

Energy efficiency (W per op/s) (=energy/op)
Performance ($ per op/s) (~~ mm^2 per op/s)

But, how do we weight these metrics?

Energy-Delay Product?
Energy-Delay Squared?

ASIC Cloud Design: Key Metrics

How do we reason about optimality?

Typical Accelerator Metrics in ISCA papers:

Energy efficiency (W per op/s) (=energy/op)
Performance ($ per op/s) (~~ mm^2 per op/s)

But, how do we weight these metrics?

Energy-Delay Product?
Energy-Delay Squared?

Datacenter TCO analysis provides the answer!!!

We include all costs for the server BOM, including silicon, DC/DC, PSU, MB, fans, …
Then we apply the Barroso et al Datacenter analysis, factoring in energy costs
Conservative assumption: 1.5 year lifetime of ASIC

Moreover, we can jointly specialize the ASIC cloud server and chip design to optimize TCO.
Observation → *Voltage scaling is a first-class optimization for TCO.*

Our Four ASIC Cloud Designs

We design ASIC Clouds for 4 application domains…

Bitcoin Mining

Litecoin Mining

These ASIC Clouds already exist ”in the wild”!

Video Transcoding (e.g. YouTube)

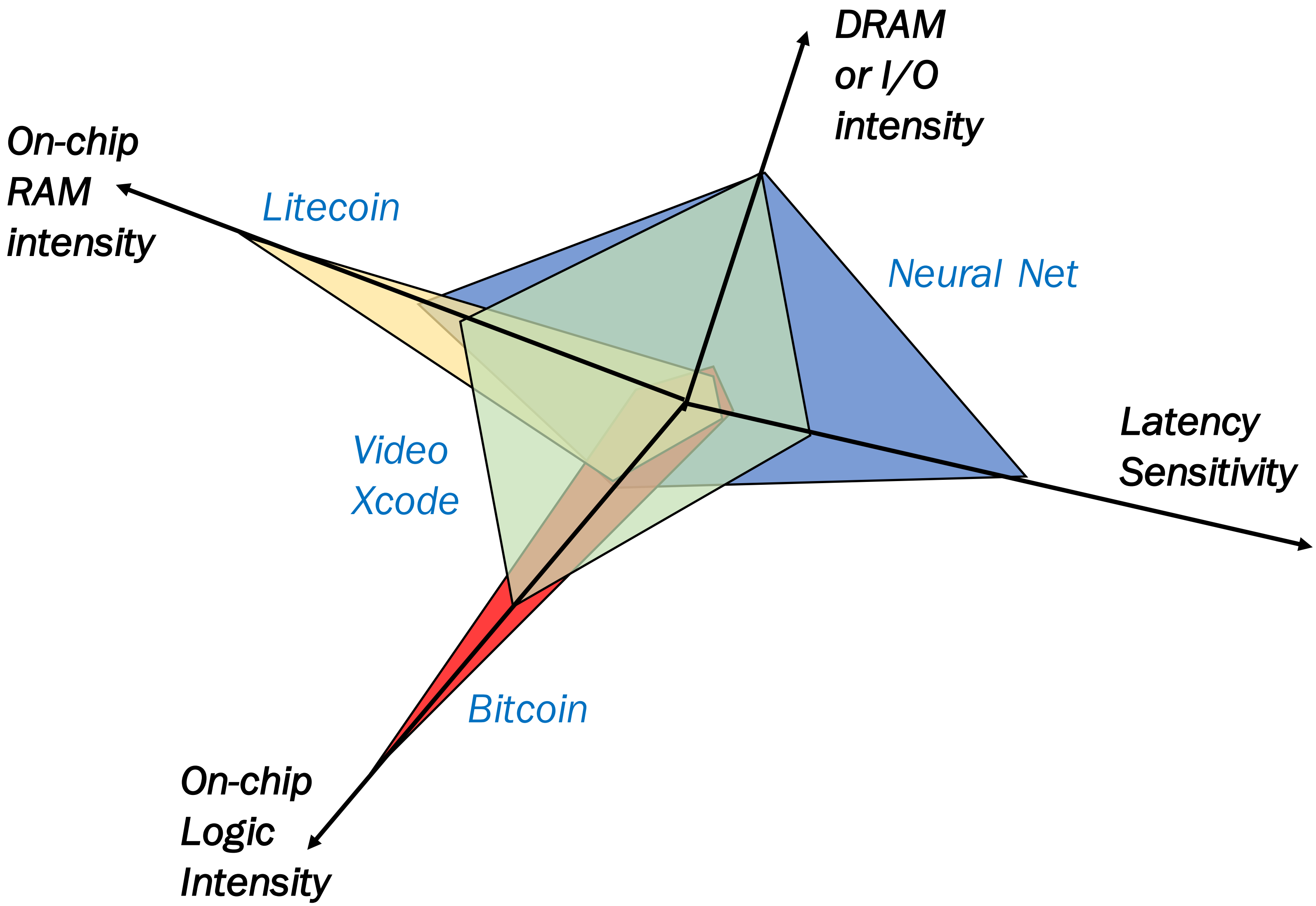
We do H.265 transcoding.

Deep Neural Networks (of course!)

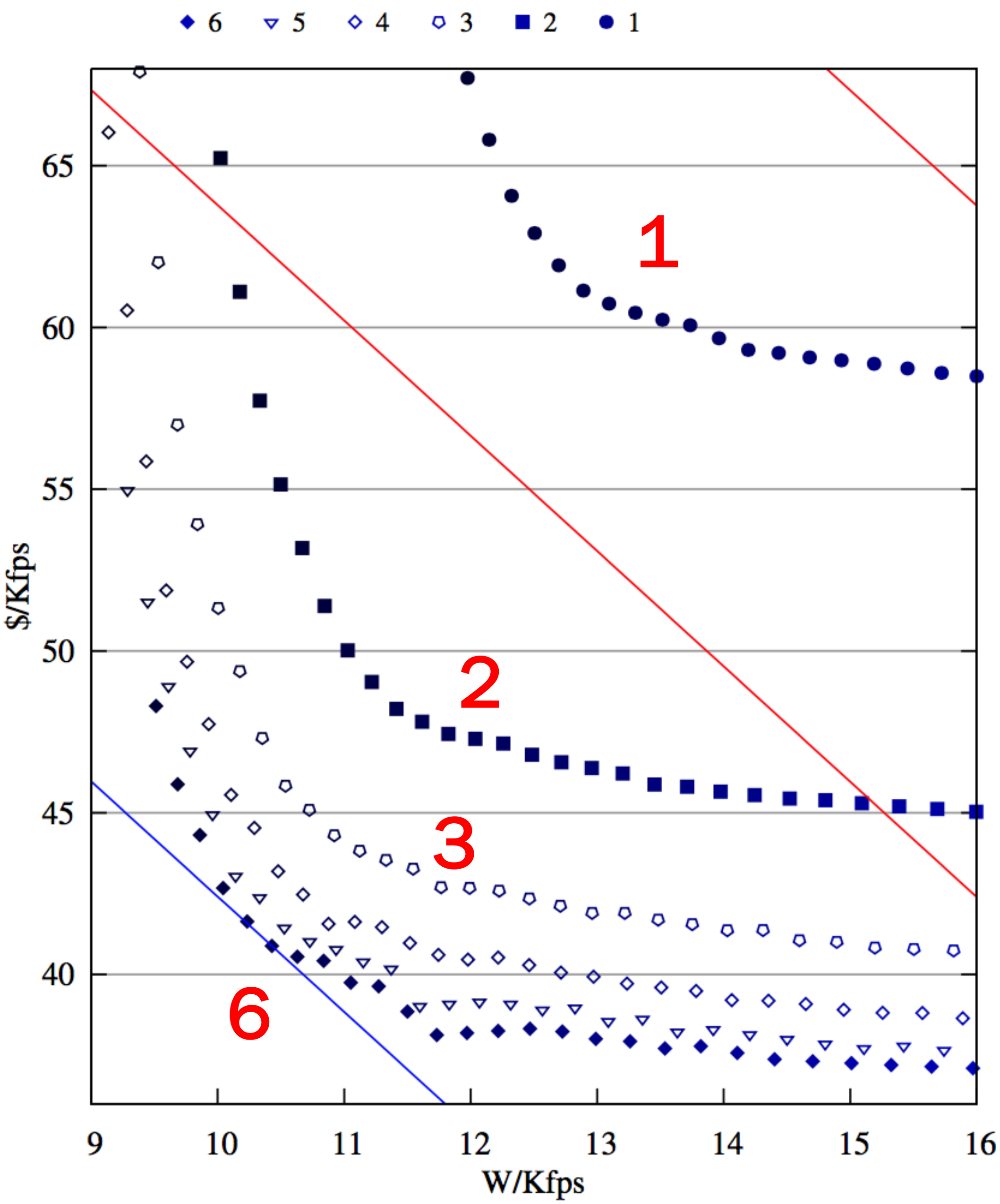
Scaling up DaDianNao into an ASIC cloud.

Accelerator Properties

We explore applications with varying properties



Video Transcoding Pareto



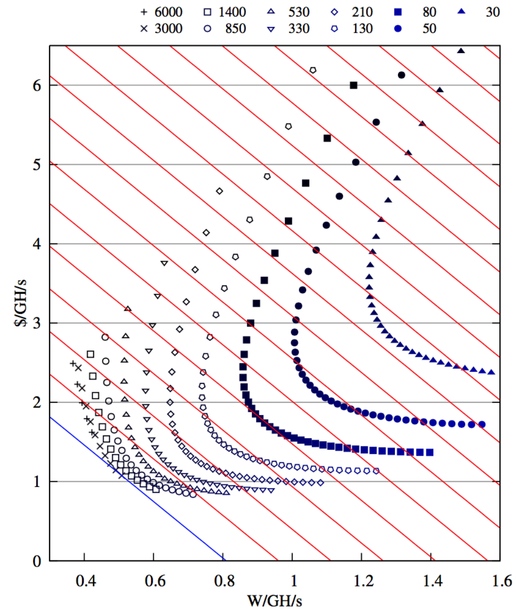
Point Series:
  # of DRAMs per ASIC

Each point in series:
  Voltage

Video Transcoding: Optimal Points

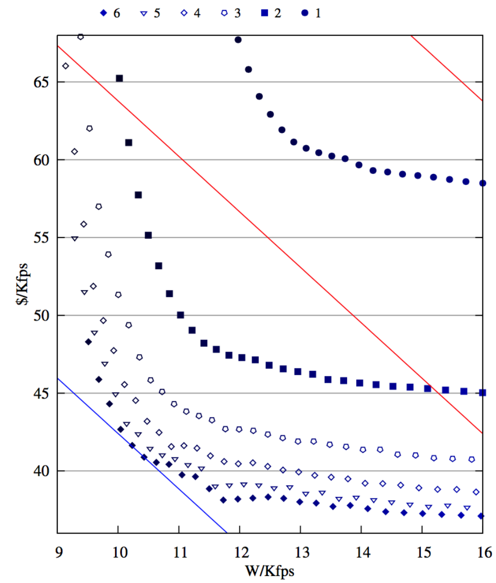
	W/Kfps Optimal	TCO/Kfps Optimal	$/Kfps Optimal
# of DRAMs per ASIC	3	6	9
# of ASICs per lane	8	5	4
# of Lanes	8	8	8
Logic Voltage (V)	0.53	0.80	1.40
Clock Frequency (MHz)	163	439	562
Die Size (mm^2)	595	456	542
Silicon/Lane (mm^2)	4,760	2,280	2,168
Total Silicon (mm^2)	38,080	18,240	17,344
Kfps/server	127	159	190
W/server	1,109	1,654	3,216
$/server	10,779	6,482	6,827
W/Kfps	8.741	10.428	16.904
$/Kfps	84.975	40.881	35.880
TCO/Kfps	129.416	86.971	107.111
Server Amort./Kfps	89.224	42.925	37.674
Amort. Interest/Kfps	5.483	2.638	2.315
DC CAPEX/Kfps	21.015	25.07	40.639
Electricity/Kfps	7.590	9.055	14.678
DC Interest/Kfps	6.105	7.283	11.806

See Paper for All Applications

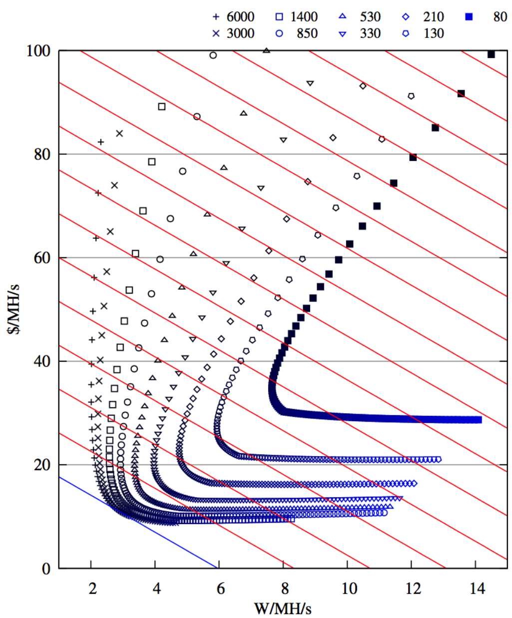


Bitcoin

Video Transcoding

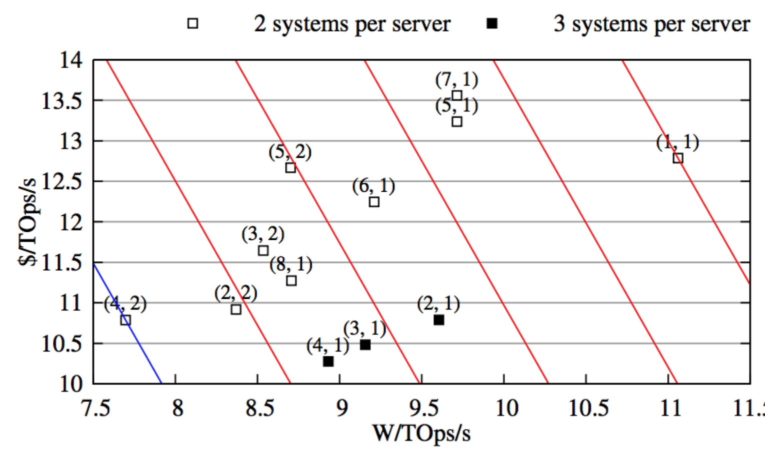


<u>TCO improvement:</u>
<u>Geomean Of 369</u>
<u>in UMC 28 nm</u>



Litecoin

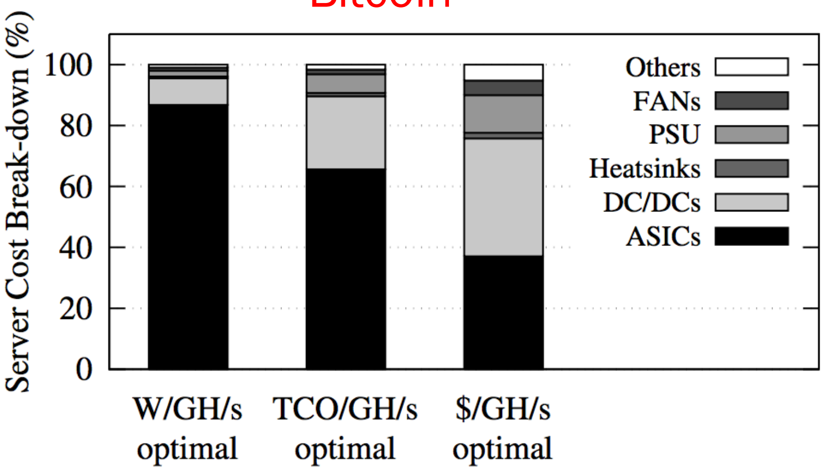
DNN



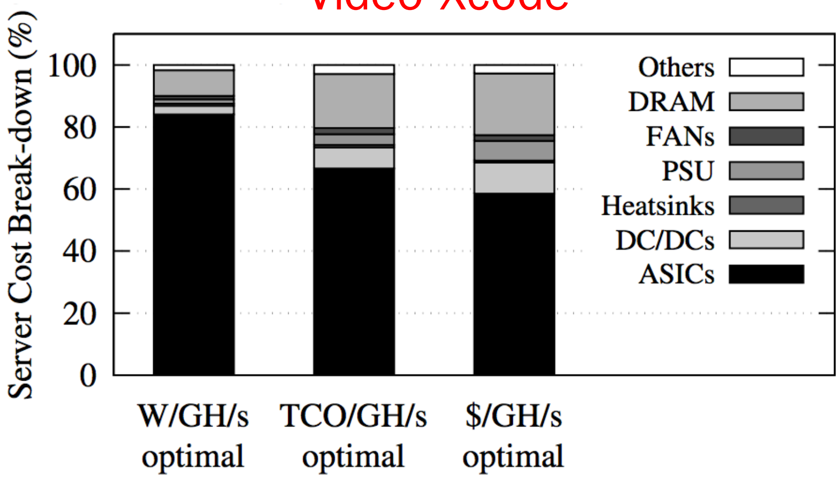
Varying silicon, chips per lane, voltage, DRAMs, RCAs per ASIC, ...

Cost Breakdowns: Two examples

Bitcoin



Video Xcode



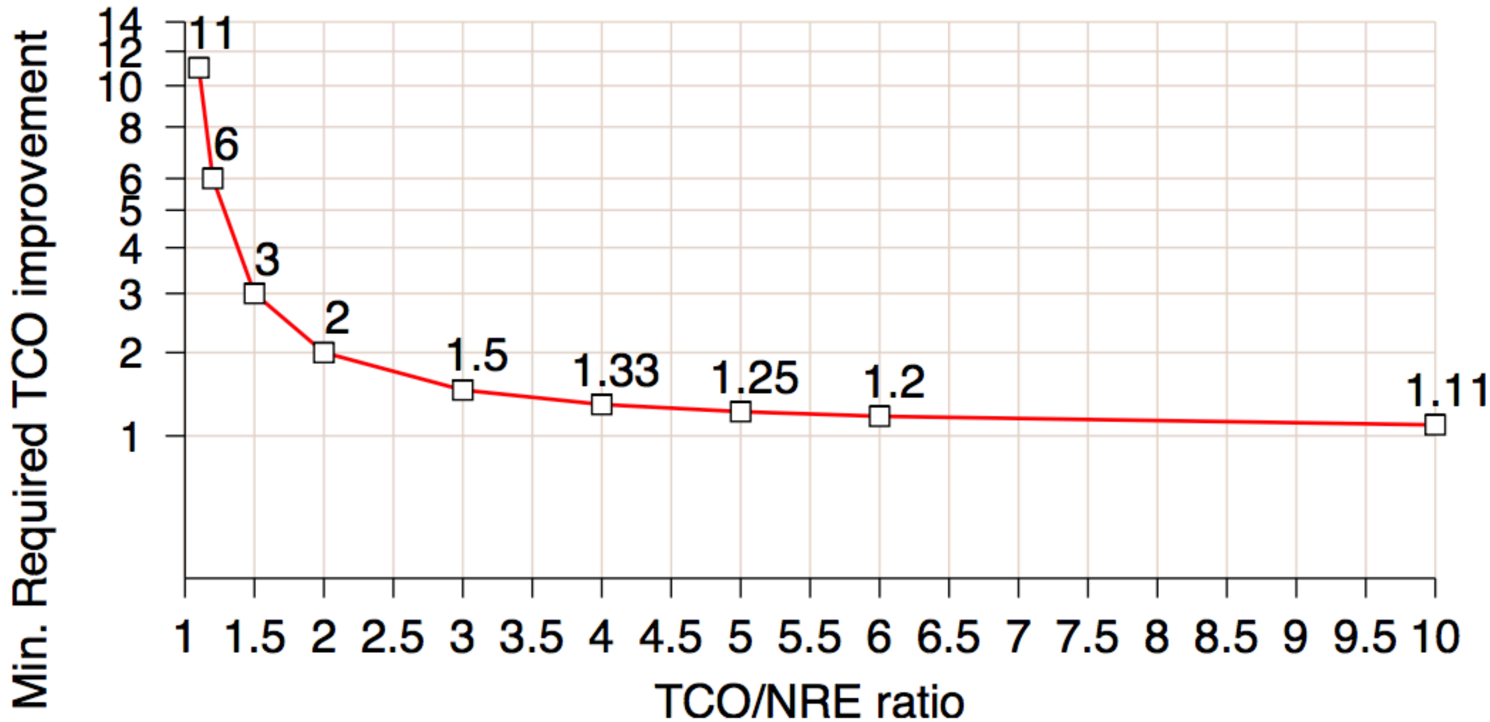
Energy optimal versions: **very low voltages** and **lots of silicon.**
Cost optimal versions: **higher voltages** and **less silicon.**
TCO optimal versions: **in between**

Bitcoin has very large DC/DC converter cost because it is so compute intensive (a "worst case" for dark silicon.)

Video Xcode instead spends extra money on DRAM.

When do we go ASIC Cloud?

When do TCO benefits outweigh ASIC development costs?



"Two-for-two" rule: If the non-ASIC TCO exceeds the ASIC NRE by 2X, and the improvement in TCO is at least 2X, then you will at least breakeven…

Interestingly, the higher your pre-ASIC TCO, the less speedup you need!

ASIC Cloud: Conclusions

ASIC Clouds are a promising direction for deploying new kinds of accelerators targeting large, chronic workloads.

We show a complete development path from Verilog to TCO-optimized ASIC Cloud datacenter.

We introduce the “two-for-two” rule, which shows that the scale of the computation affects how much speedup you need to merit going ASIC Cloud.

*****