Moonwalk:
NRE Optimization in ASIC Clouds

or, accelerators will use old silicon

Moein Khazraee, Lu Zhang, Luis Vega,
and Michael Bedford Taylor
UC San Diego

Compute trends in 2017

- Bifurcation of computation into Client and Cloud
  - Client is mobile SoC
  - Cloud is implemented by datacenters

Compute trends in 2017

- 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] (power wall)
  - 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 2017

- 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] (power wall)
  - 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.

Early Signs of Specialization in the Datacenter

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

Early Signs of Specialization in the Datacenter

- Xeon Processors
  - Xeon-D [Facebook]
  - Customer specialized SKUs [Oracle]
- GPU-based clouds
  - Deep Neural Networks [Baidu Minwa]

Early Signs of Specialization in the Datacenter

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

Early Signs of Specialization in the Datacenter

- Xeon Processors
  - Xeon-D [Facebook]
  - Customer specialized SKUs [Oracle]
- GPU-based clouds
  - Deep Neural Networks [Baidu Minwa]
- FPGA-based clouds
  - Catapult [Microsoft]
  - High Frequency Trading [Most Wall Street firms]
- What about ASIC-based clouds?

ASIC Clouds: Key Motivation

- The Cloud model leads to growing classes of planet-scale computations
  - Facebook runs face recognition on 2B pics/day
  - Siri recognizes speech for ~1 Billion iOS user
  - YouTube Video Transcodes to Google VP9 for the 500 hours uploads per minute

- These computations incur high Total Cost of Ownership (TCO) for the provider

ASIC Clouds: Key Motivation

- These cloud computations are *scale-out*:
  we are doing the same computation across millions or billions of users

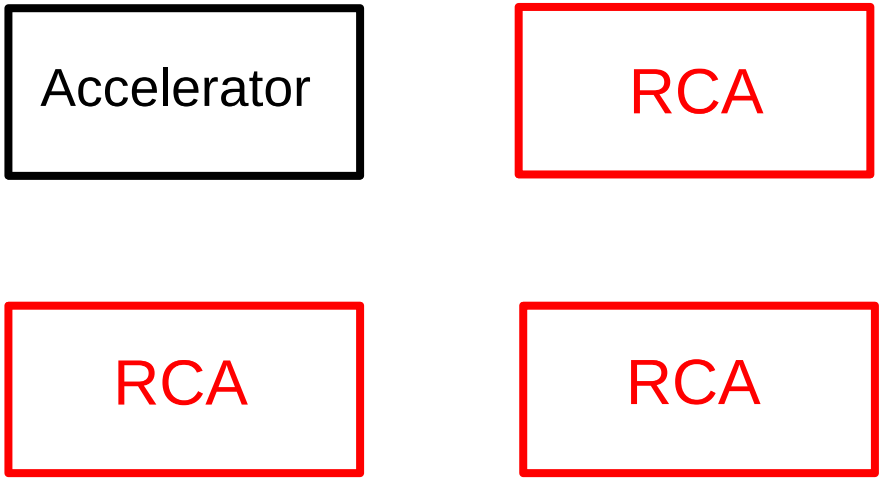
- As these computations become sufficiently large, we can specialize the hardware for that particular computation to reduce TCO.

- Lowering Non-Recurring Engineering cost (NRE) is a key factor for ASIC cloud feasibility.
  - Our paper makes a key contribution by showing how to calculate NRE for an ASIC Clouds.

ASIC Cloud Architecture

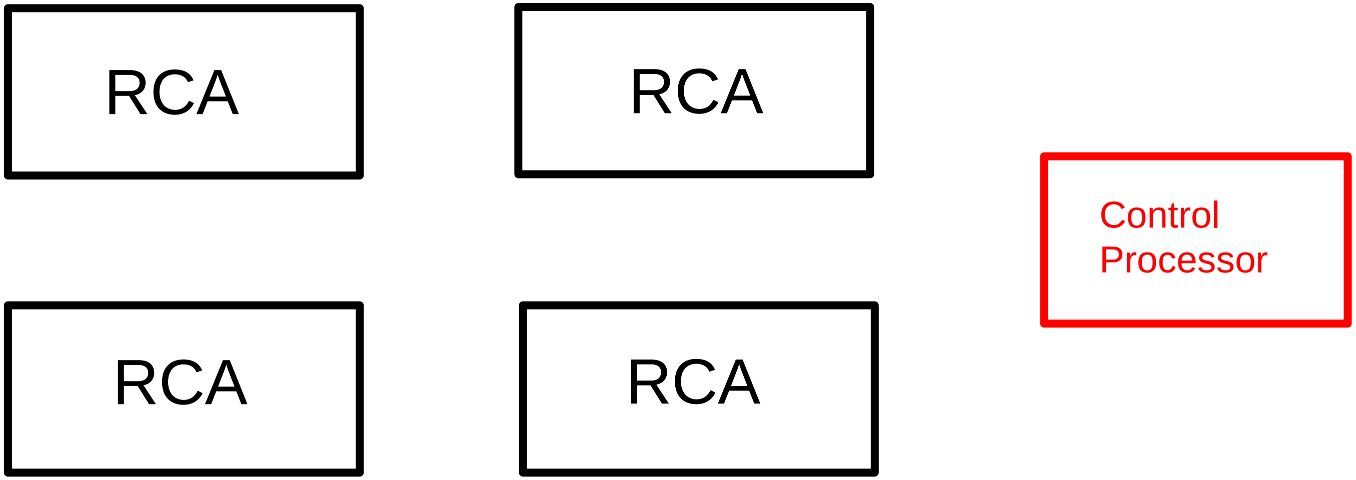
It all starts with an accelerator for a planet-scale computation. Maybe it’s a commercial IP core, or custom designed widget in Verilog.

ASIC Cloud Architecture



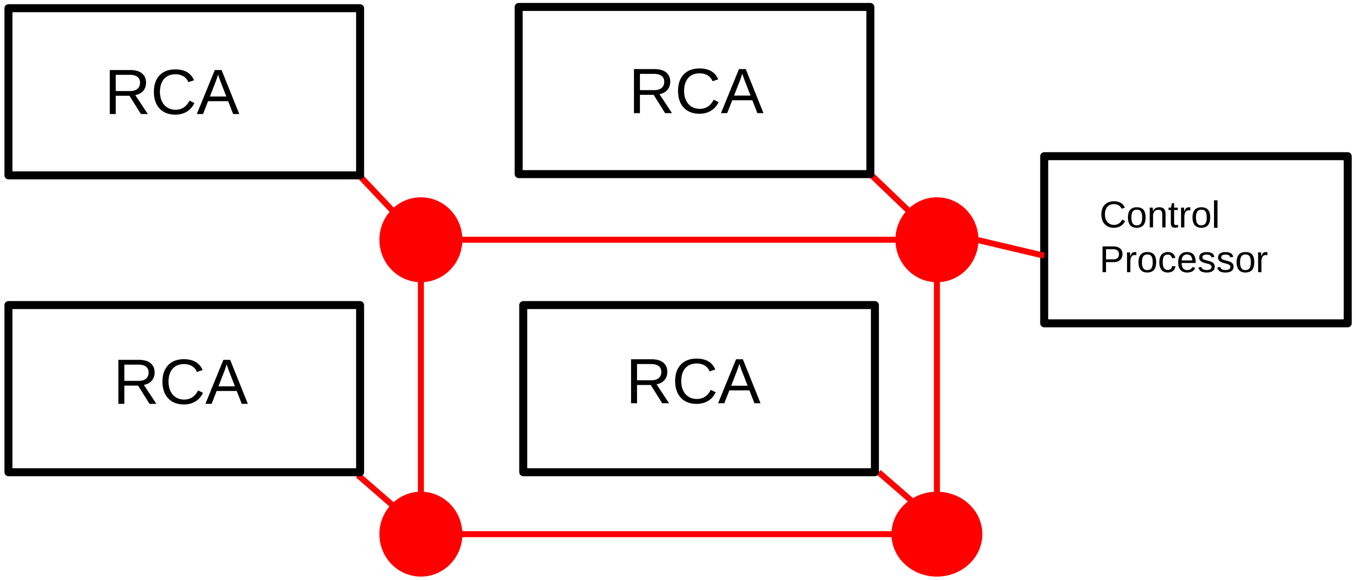
Replicate this accelerator multiple times inside an ASIC die. We’ll now call it a “replicate compute accelerator”, or “RCA”.

ASIC Cloud Architecture



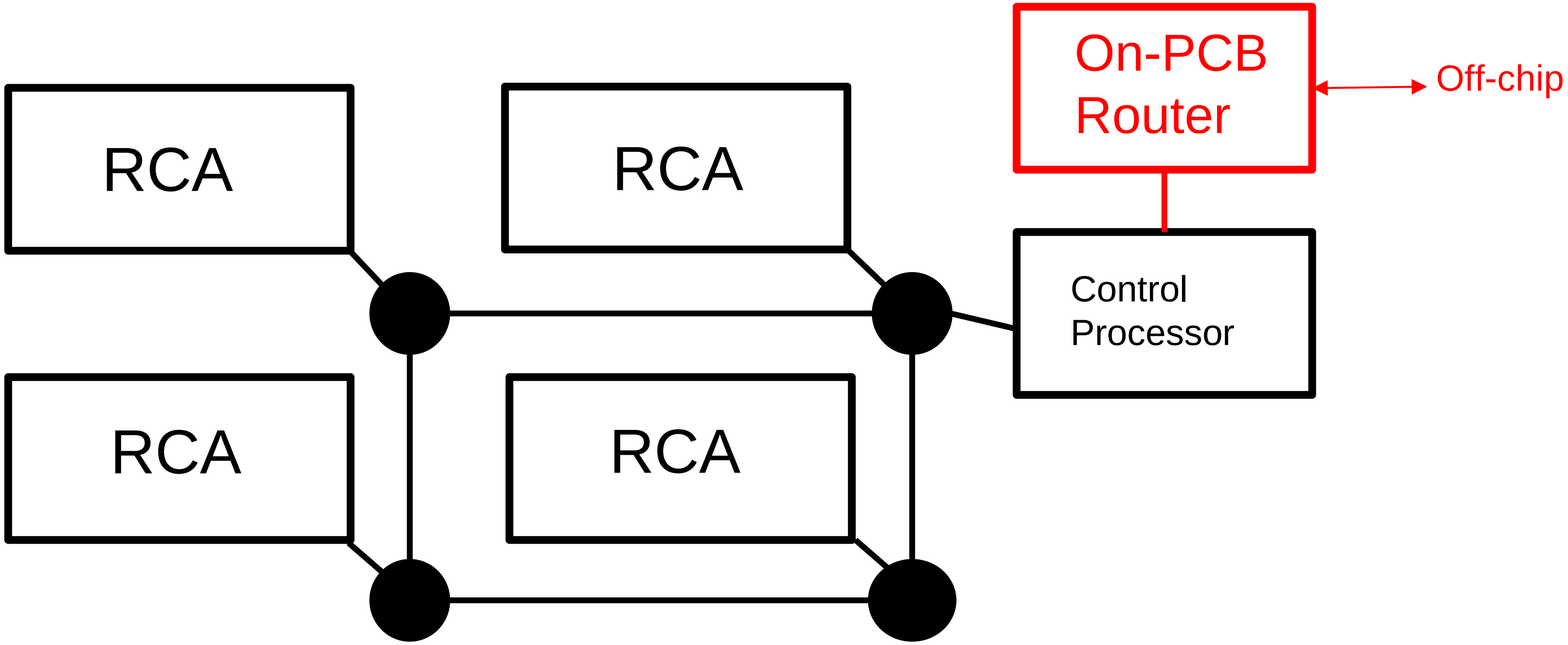
Then we add a control processor to distribute work and schedule computation onto the RCAs.

ASIC Cloud Architecture



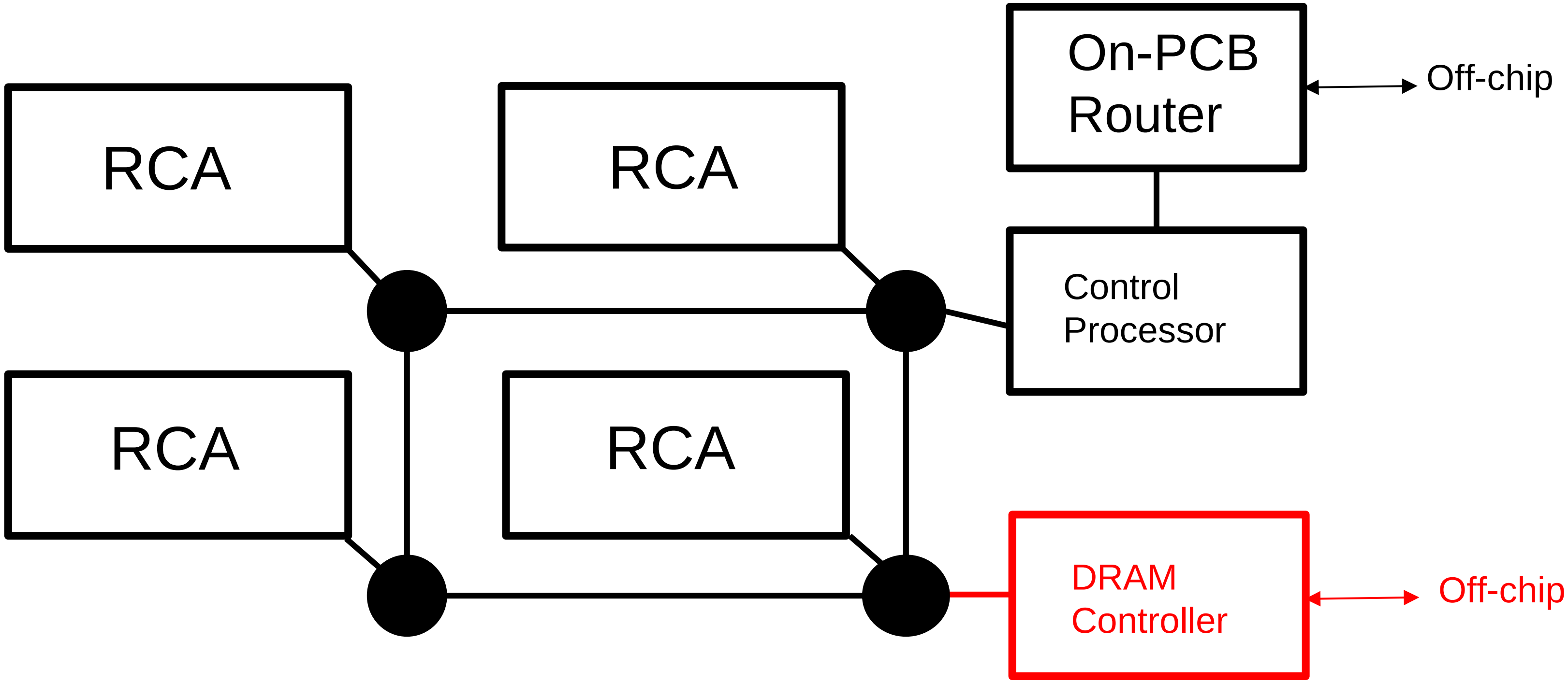
Work is distributed over a very simple on-chip network, the **On-ASIC Network**, which is provisioned according to the needs of the RCAs. RCA's usually do not talk to each other.

ASIC Cloud Architecture



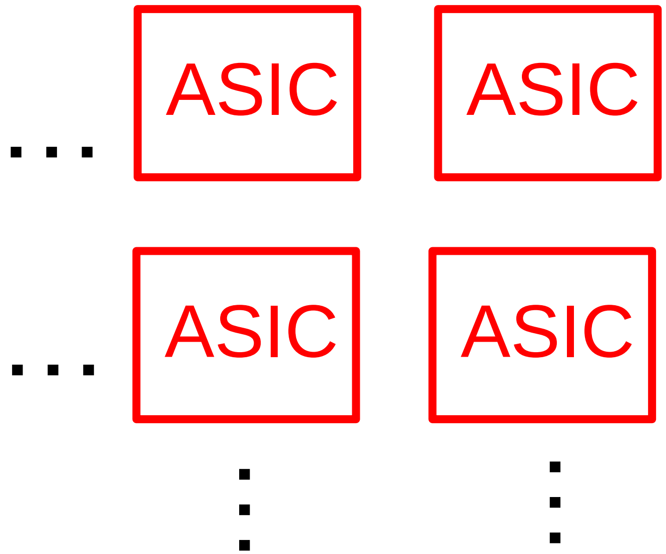
The control processor receives work from off-chip via the On-PCB router.

ASIC Cloud Architecture



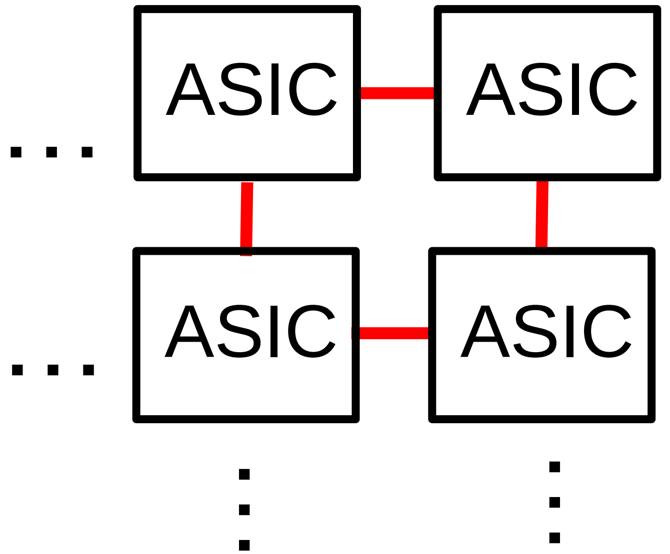
For those accelerators that need off-chip DRAM, we add shared DRAM controllers. Finally bake it into an ASIC: PLL, Clock Tree, Power Grid, Flip Chip BGA Packaging...

ASIC Cloud Architecture



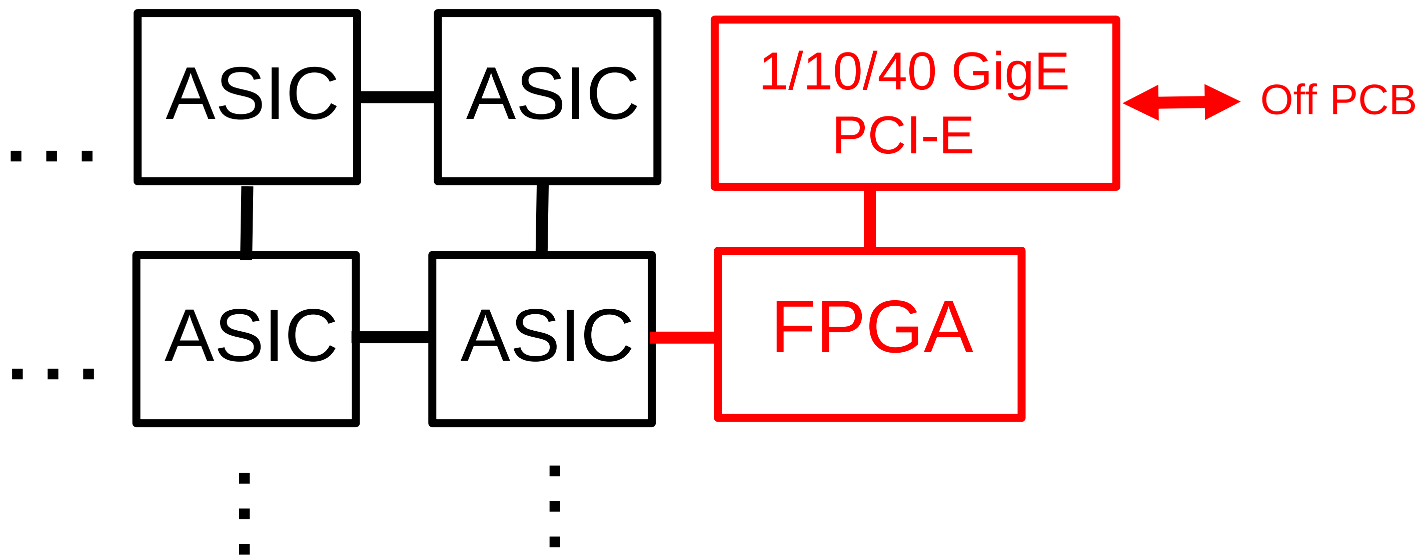
Then build the PCB by replicating ASICs across the PCB

ASIC Cloud Architecture



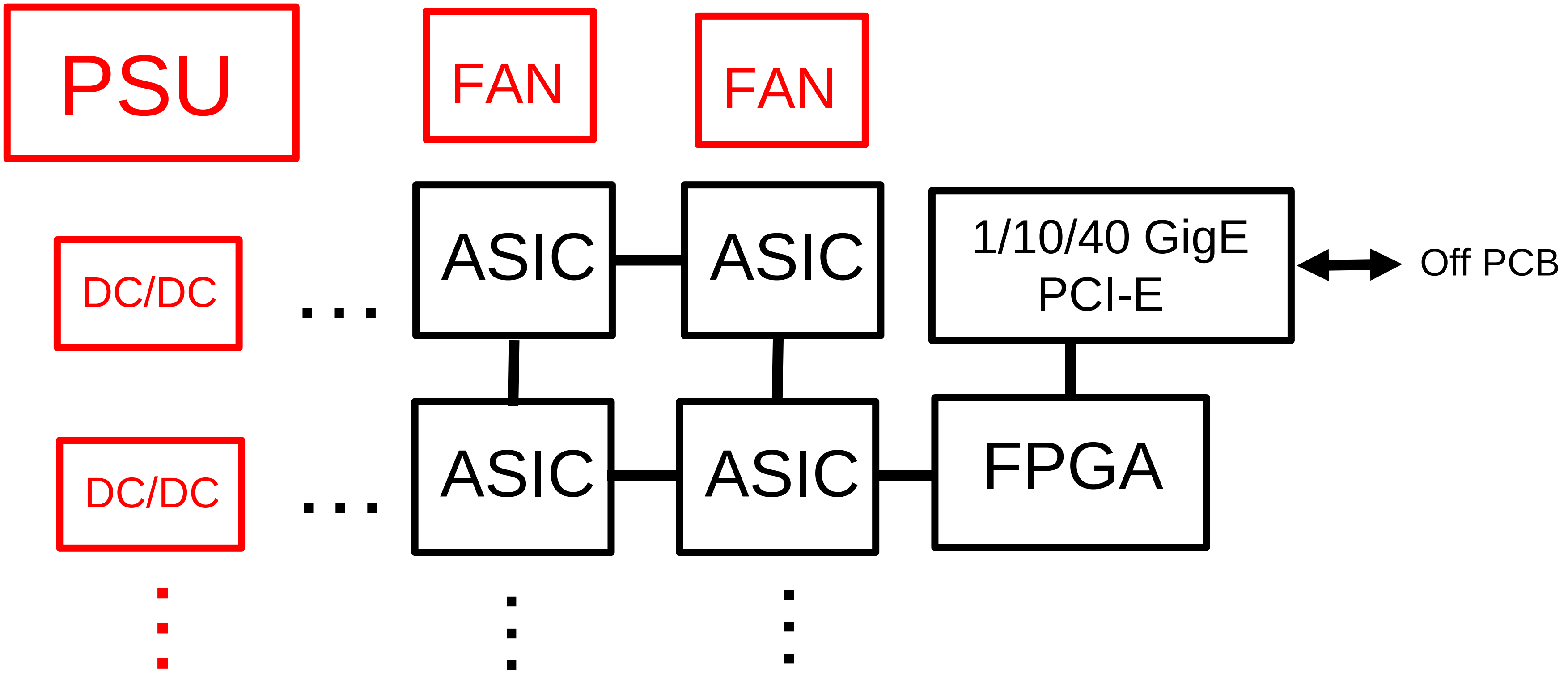
Connect their on-PCB routers via PCB traces

ASIC Cloud Architecture



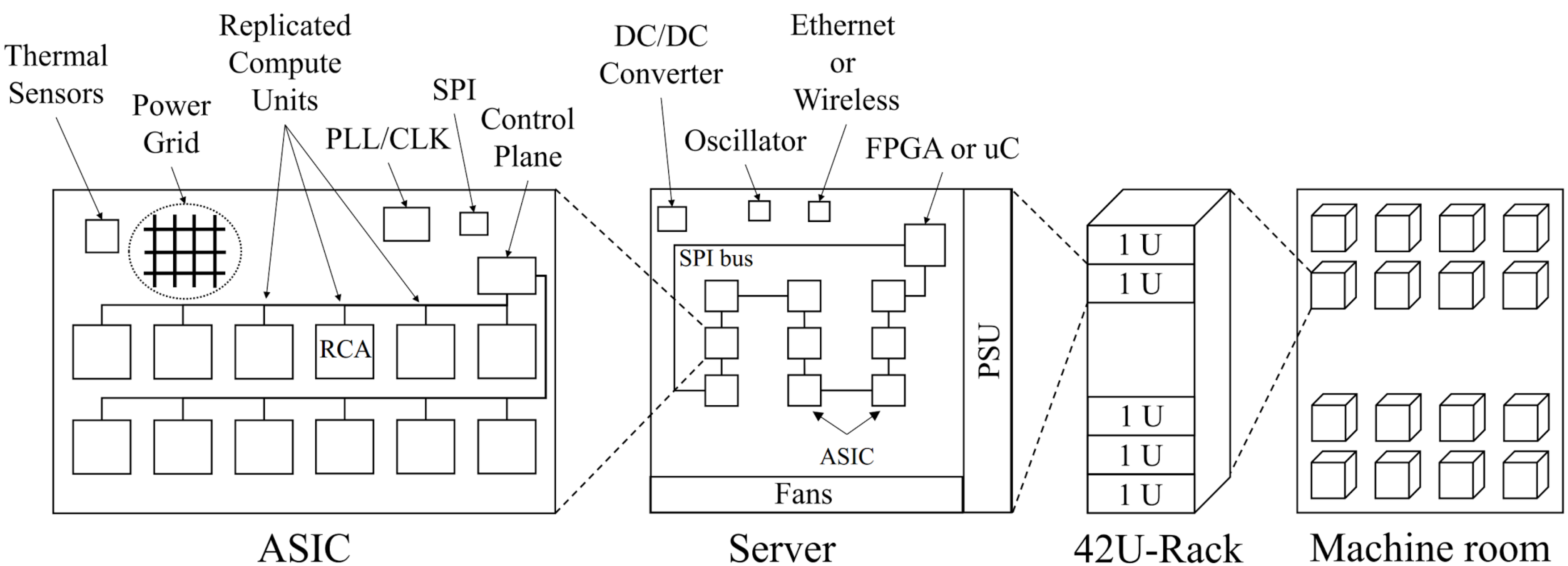
Connect the on-PCB network to an FPGA that routes data from off-PCB interface (e.g. GigE, PCI-E)

ASIC Cloud Architecture



Then we add the plumbing: DC/DC, Fans, Heatsinks and PSU. The PCB goes inside the chassis and we have an ASIC cloud server.

ASIC Cloud Architecture



- 1U servers are packed into standard 42U racks.
- Racks are integrated into machine room.

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 (face/voice recognition)
  - Scaling up DaDianNao into an ASIC cloud.

ASIC Cloud Design: Key Metrics

- Accelerator Metrics:
  - Energy efficiency (W per op/s) (=energy/op)
  - Performance ($ per op/s)

- Conventional trivial weighing:
  - Energy-Delay product or Energy-Delay squared

- Datacenter Total Cost of Ownership as the new metric
  - Barroso et al Datacenter analysis
  - Conservative assumption: 1.5 year lifetime of ASIC

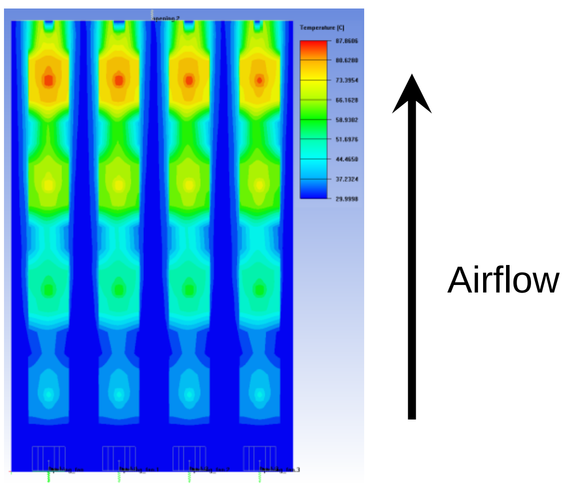
Barroso et al, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," Synthesis lectures on computer architecture, vol. 8, 2013

Complete Design Methodology from Verilog to TCO-Optimized Datacenter

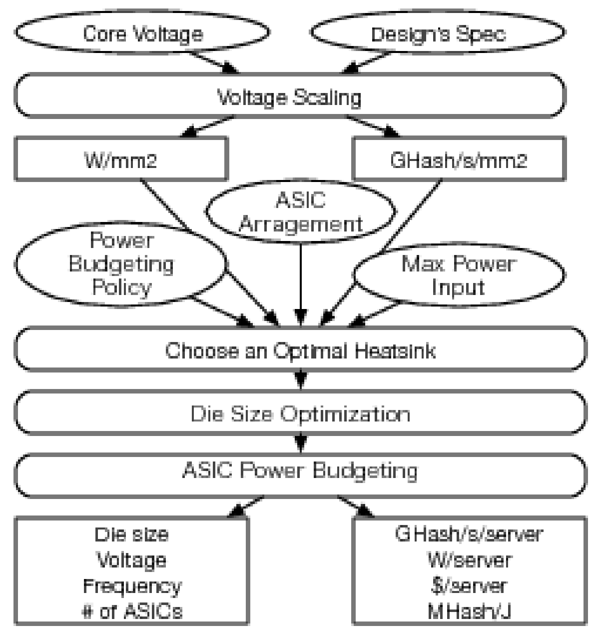
- We can jointly specialize server and ASIC to optimize TCO.
- Thermal optimization based on RCA properties:

  ASIC placement (DUCT layout), heat sink optimization (# fins, width, materials and depth), die size

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



Complete Thermal Analysis using CFD (Ansys ICEpak)

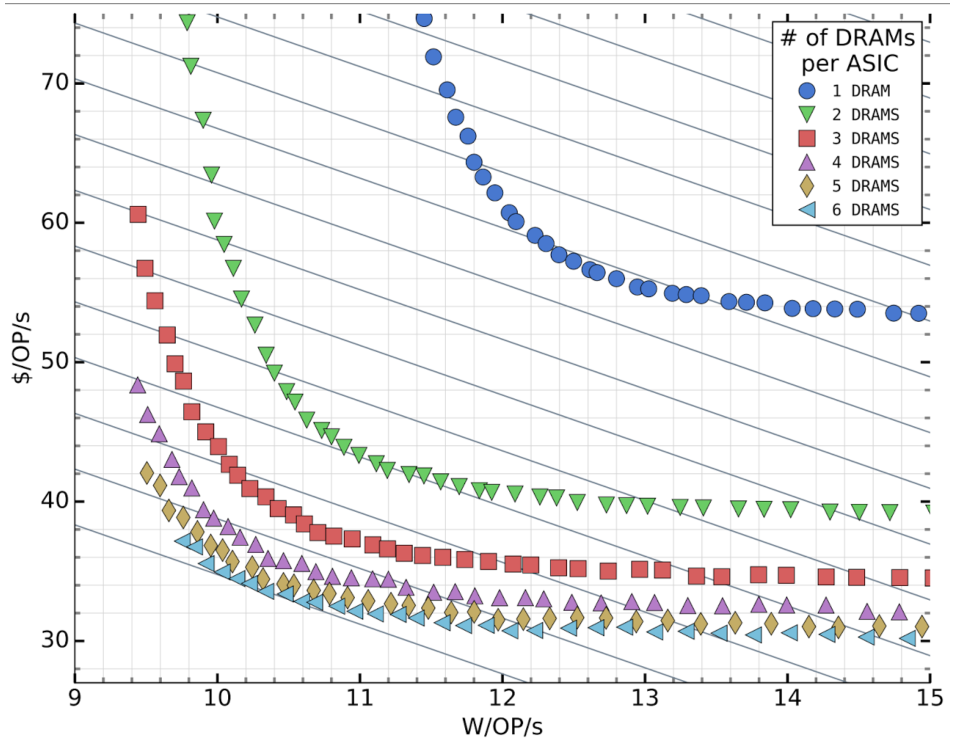


A flow that converts ASIC properties to Server properties and TCO

Design Space Exploration

- Observation: Voltage scaling is a first-class optimization for TCO. Core voltage increases from left to right
- Pareto curve for Performance and energy efficiency
- Diagonal lines show equal TCO

– Exploring different # of DRAMs per ASIC, # of ASICs per lane, and Logic Voltage, as well as thermal optimizations



Video Transcode Pareto in 28nm

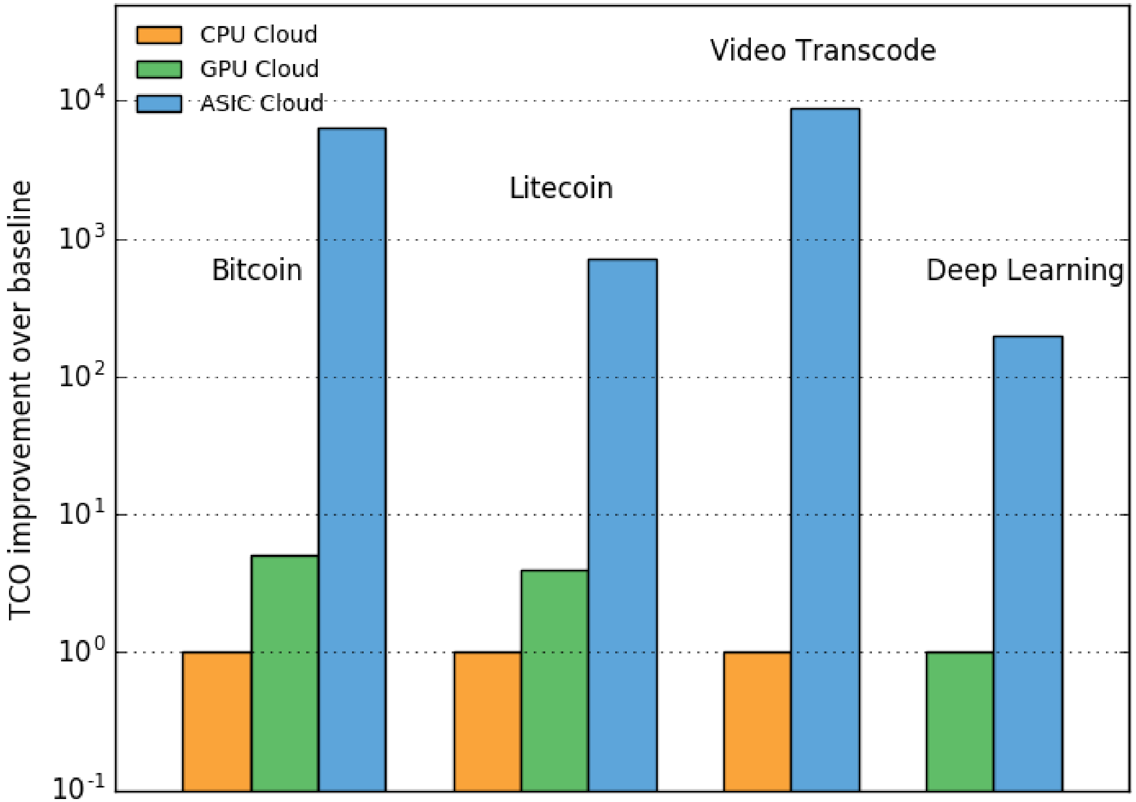
This plot is for 5 ASICs per lane.

Deathmatch

- ASIC Servers greatly outperform the best non-ASIC alternative in terms of TCO per op/s.

GPU:
- AMD 7970 for BC and LC
- NVIDIA Tesla K20X for Deep Learning

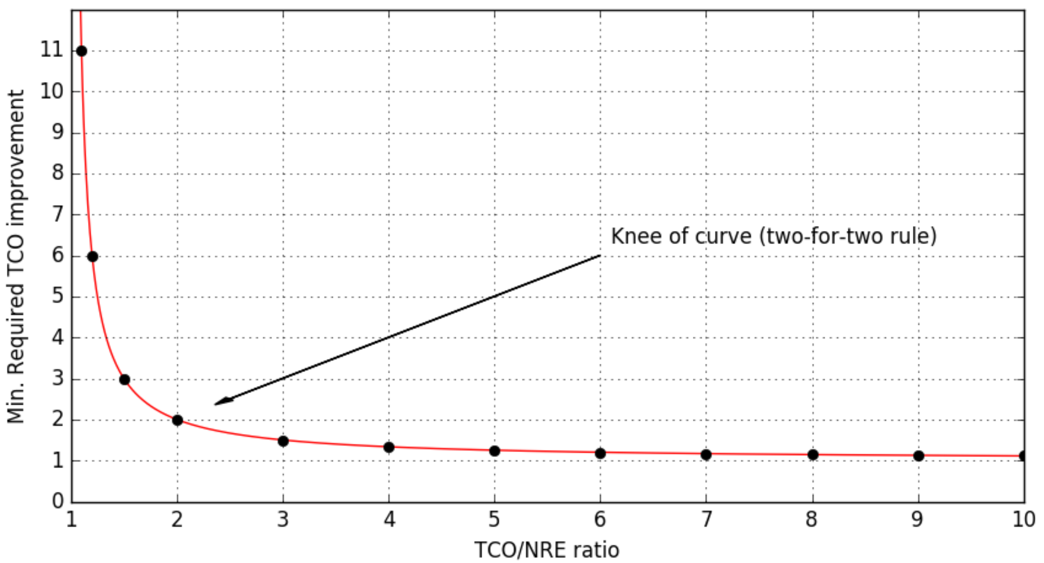
CPU:
- Core i7 3930K for BC and LC
- Core i7 4790K for Video Transcode



CPU Cloud vs. GPU Cloud vs. 28nm ASIC Cloud Deathmatch.

When do we go ASIC Cloud?

- TCO improvement vs. TCO/NRE
  - TCO improvement: determined by accelerator improvements versus best alternative
  - TCO: determined by scale of computation (higher is better)
  - NRE: determined by ASIC development and deployment costs (lower is better)
- "Two-for-two" rule: Moderate speed-up with low NRE beats high speed-up at high NRE

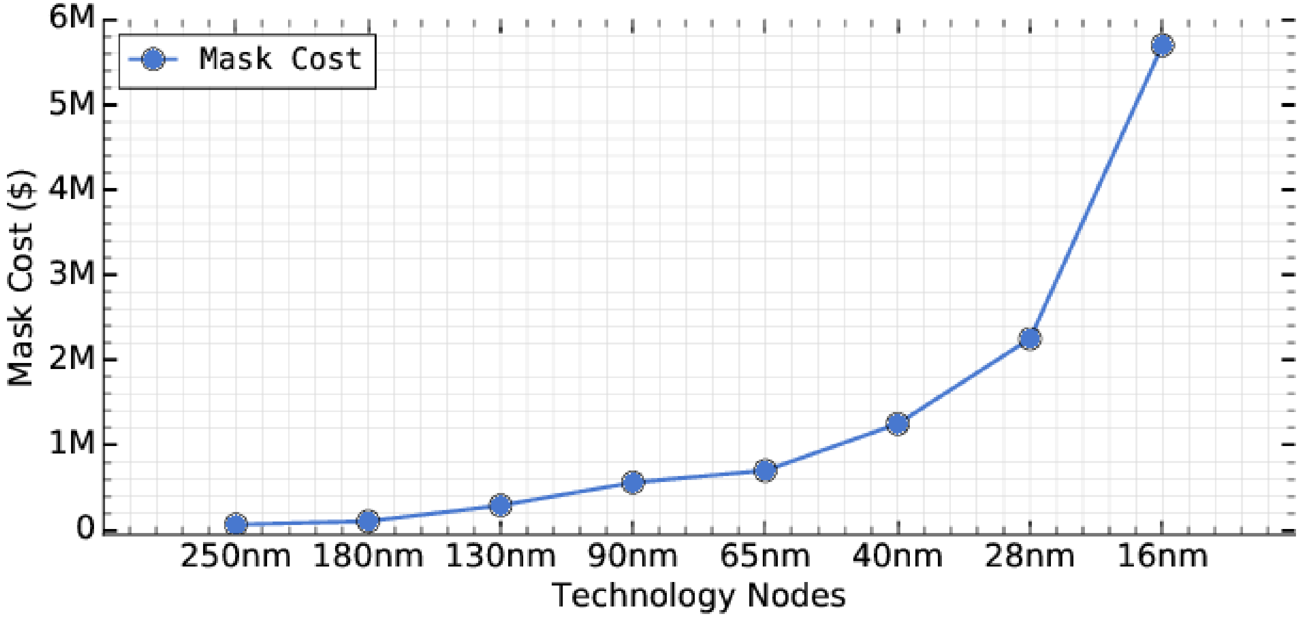


Building a model for NRE

- Mask cost
- IP licensing cost
- Labor cost (Frontend, Backend and system NRE)
- Tools cost (Frontend and Backend)
- Package NRE

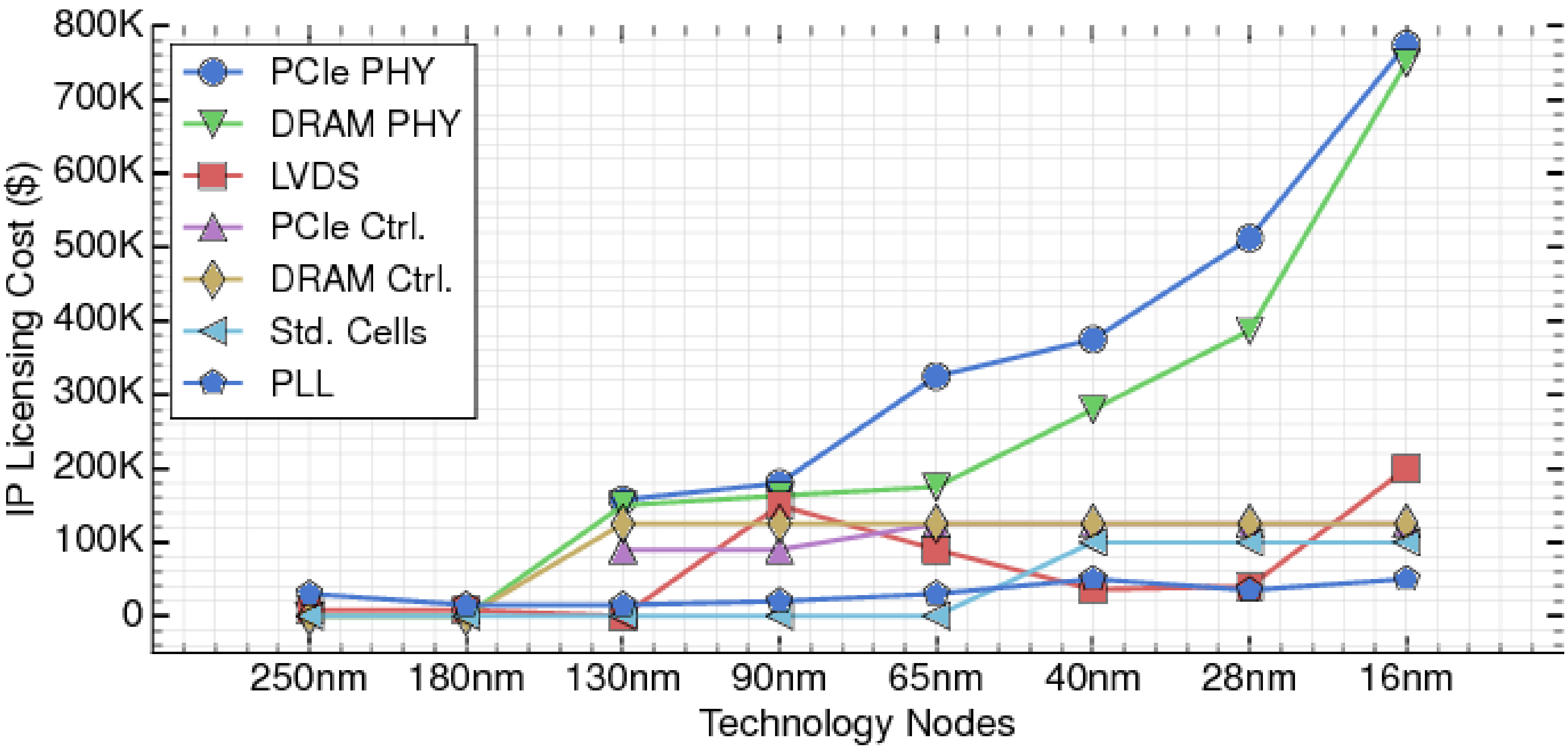
NRE: Mask and Packaging

- Mask costs rise exponentially (total 89x range)



- Package NRE is fixed among process technologies.
  – Flip-chip BGA package NRE is $105K

NRE: IP Licensing Cost



NRE: Backend Labor Cost

- Cost of pushing Verilog netlist through backend flow is fairly steady among nodes
  - But increases dramatically in double-patterned technologies like 16nm.

Tech	250nm	180nm	130nm	90nm	65nm	40nm	28nm	16nm
Backend labor cost per gate ($) [30]	0.127	0.127	0.127	0.127	0.127	0.129	0.131	0.263

H. Jones. Strategies in Optimizing Market Positions for Semiconductor Vendors Based on IP Leverage. IBS White Paper, 2014.

NRE: Labor and Tool Costs

- Backend labor time is calculated based on backend labor cost per gate model

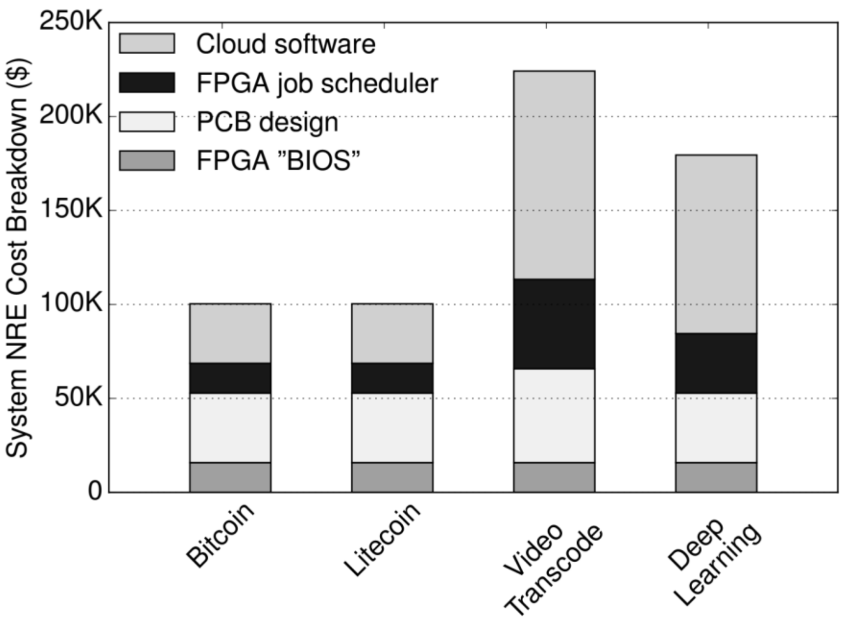
Frontend Labor Salary [19]	$/yr	115K
Frontend CAD Licenses	$/Mm	4K
Backend Labor Salary [19]	$/yr	95K
Backend CAD Licenses	$/month	20K
Overhead on Salary		65%

Values are for San Diego, 2016

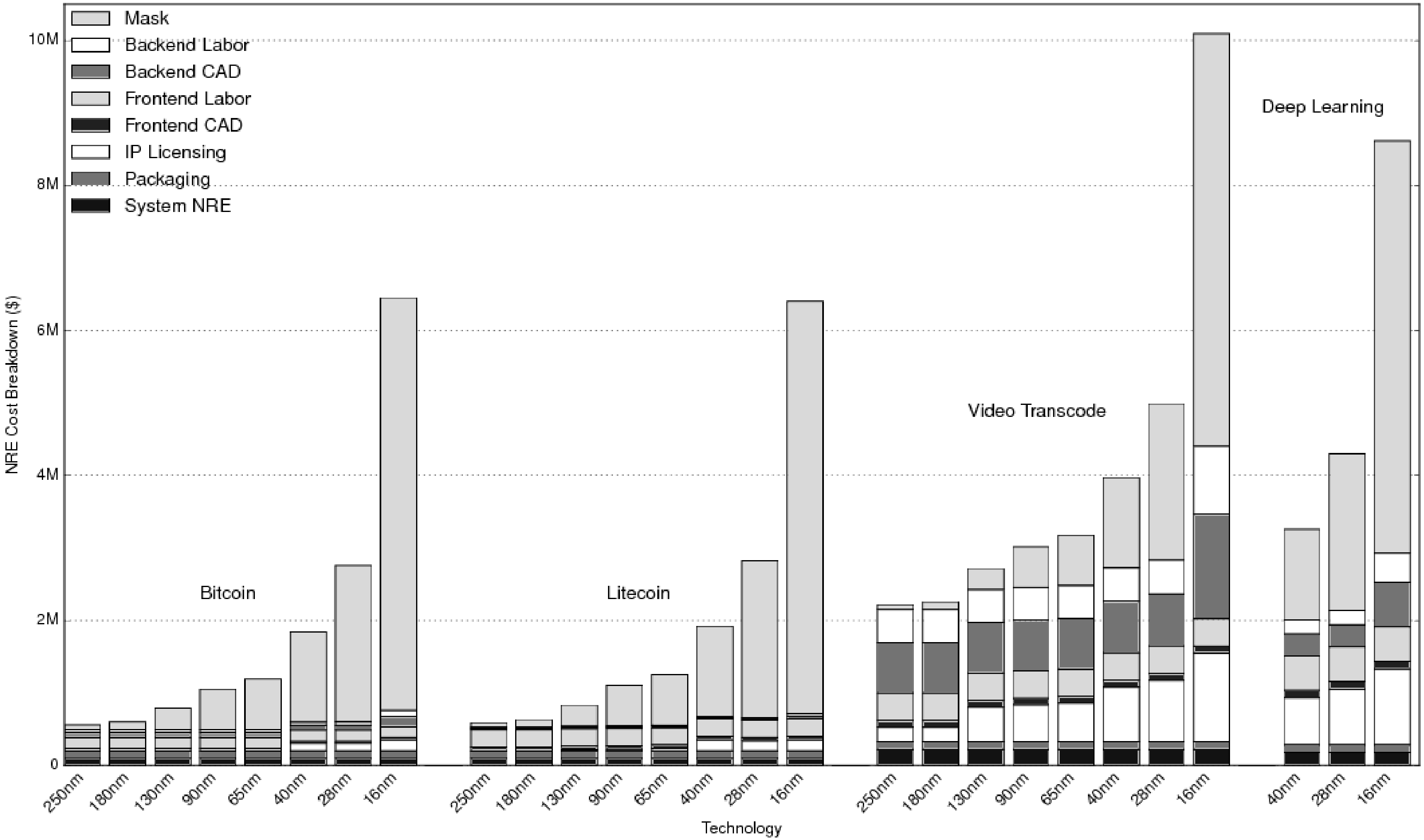
NRE: App dependent components

Application	Bit-coin	Lite-coin	Video Tr-anscode	Deep L-earning
RCA gate count	323K	96.7K	3.56M	1.51M
FE CAD-months	8	12	23	26
FE Mm	9.5	15	24	30
FPGA job distr. code, Mm	1	1	3	2
FPGA "BIOS" code, Mm	1	1	1	1
Cloud Software, Mm	2	2	7	6
PCB Design cost ($)	37K	37K	50K	37K

PCB design costs are for late 2016

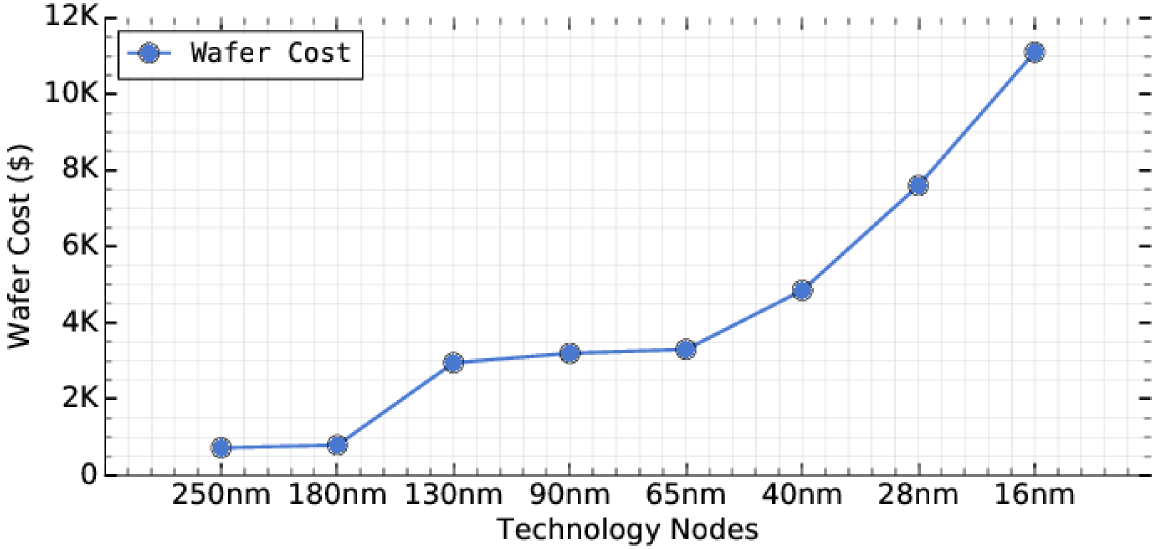


NRE breakdown for Benchmarks



Marginal Cost: Wafer and Package cost

- Wafer costs rise exponentially after 65nm; jump on transition to bigger wafers



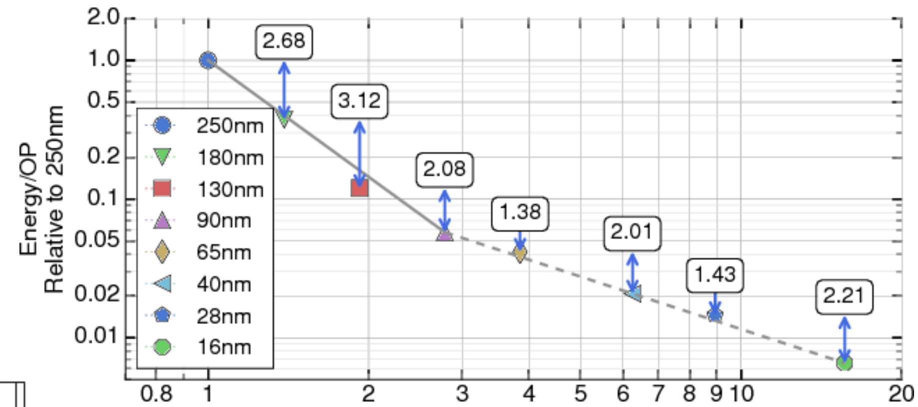
- Wafer Diameter is 200 mm until 180nm and 300 mm afterwards

Process Node as a Knob

- ASIC process technology nodes from 250 nm to 16 nm give us a range of:
  - **256**x in **maximum accelerator size**
  - **15.5**x in **max transistor frequency**
  - **152**x in **energy per op**
  - **28**x in **cost per op/s**
  - **89**x in **mask costs**

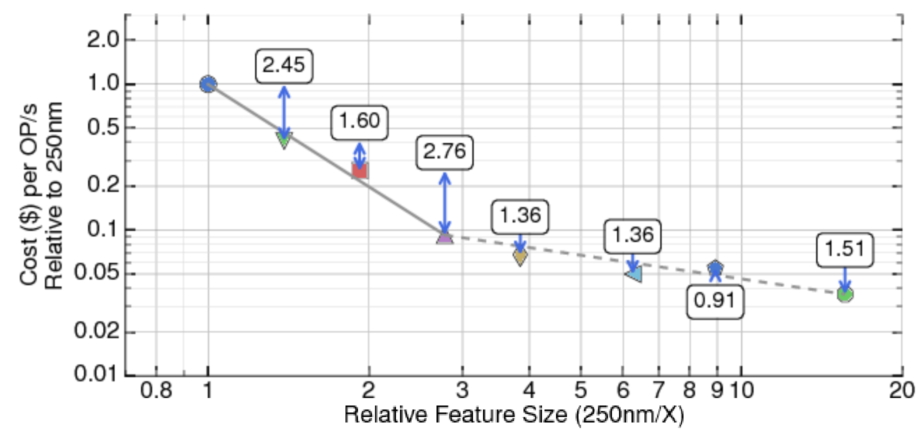
Process Node as a Knob (cont'd)

- Energy per op improvement
  - But flattens because of end of Dennard Scaling.



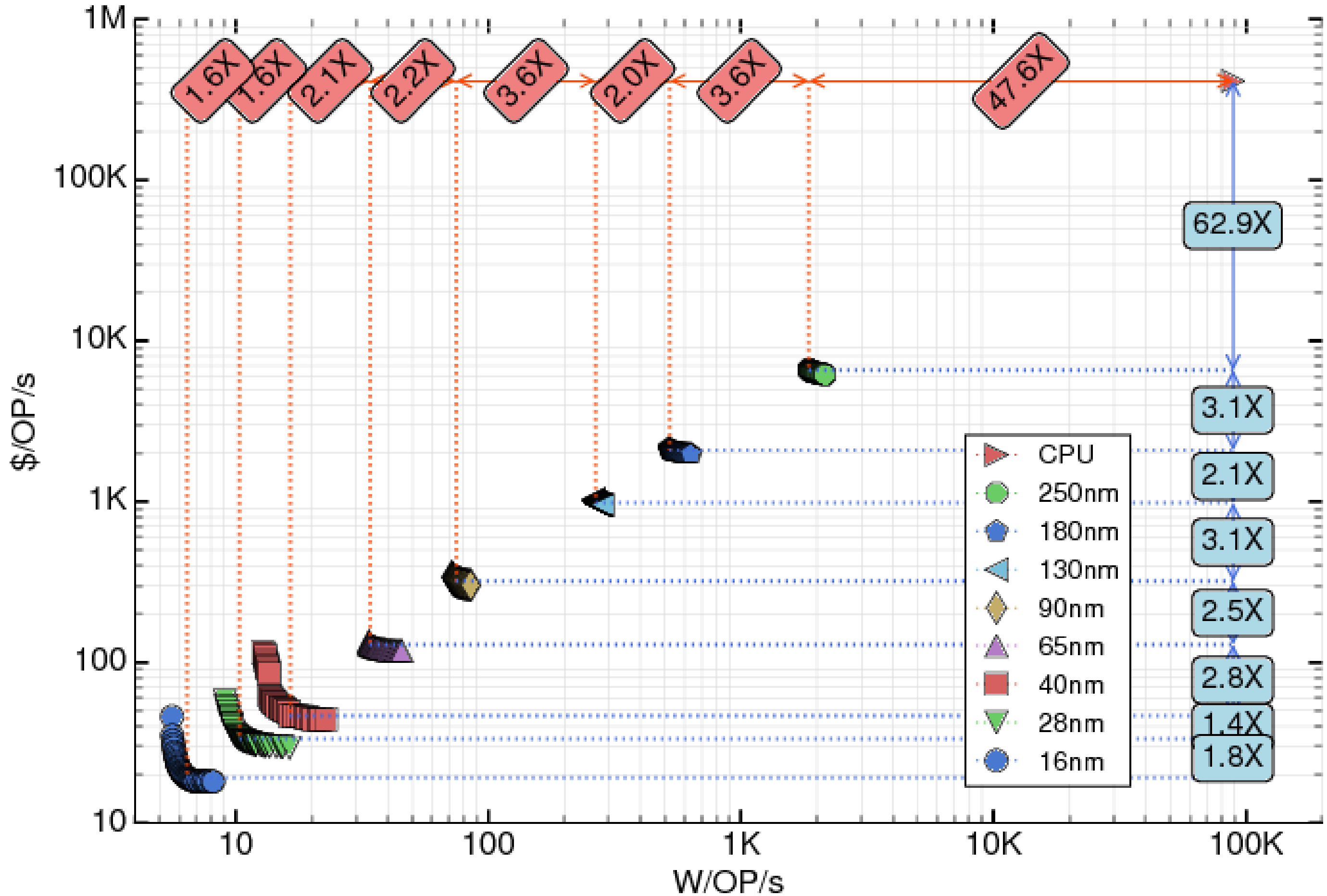
Tech Node (nm)	250	180	130	90	65	40	28	16
Nom. V_{dd} (V)	2.5	1.8	1.2	1.0	1.0	0.9	0.9	0.8

- Cost per op/s improvement
  - But flattens because of post-Dennard power density limitations and increase in wafer cost



Pareto Frontiers across technology nodes

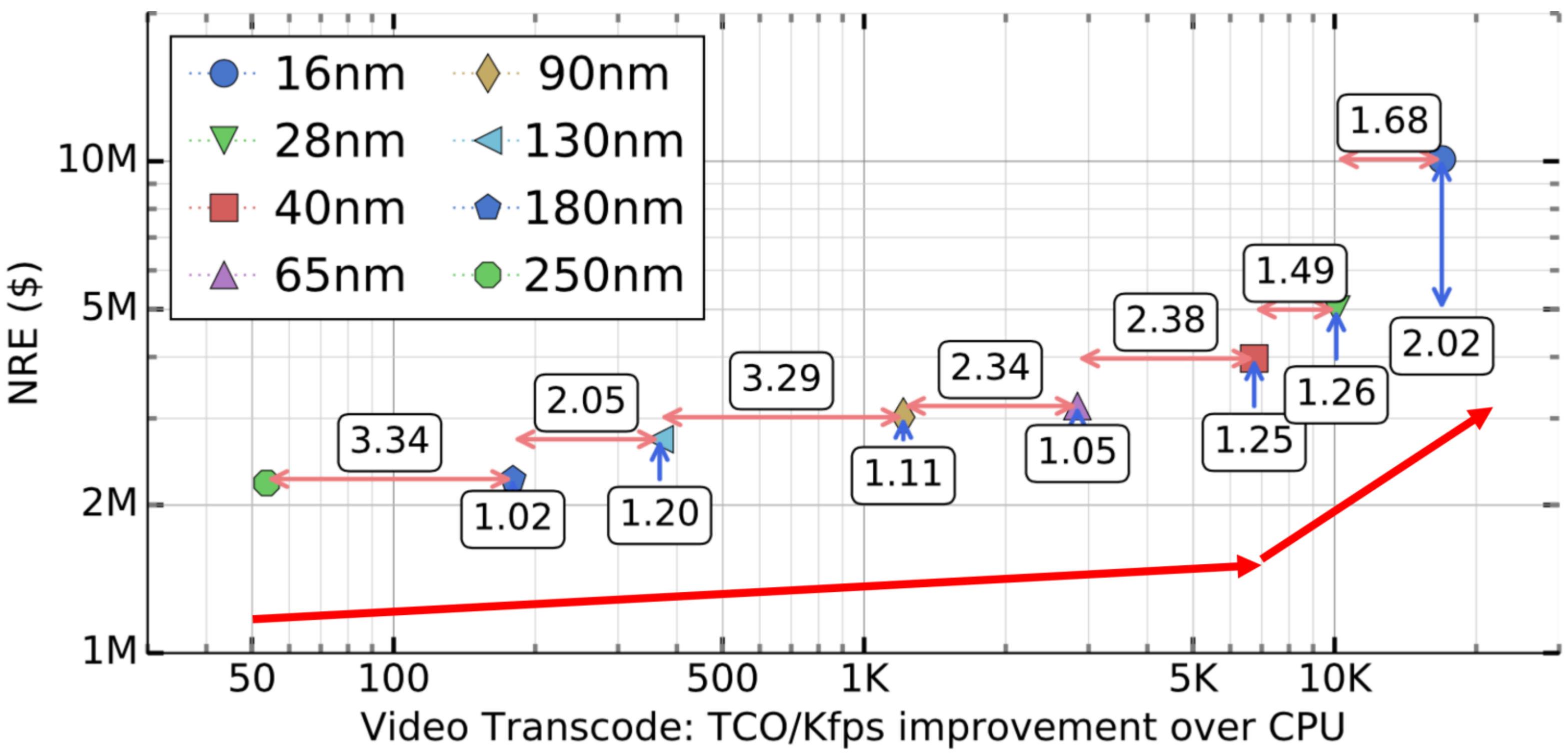
But which node is good enough?



Video Transcode Pareto frontiers improve in both energy and cost efficiency for newer technologies.

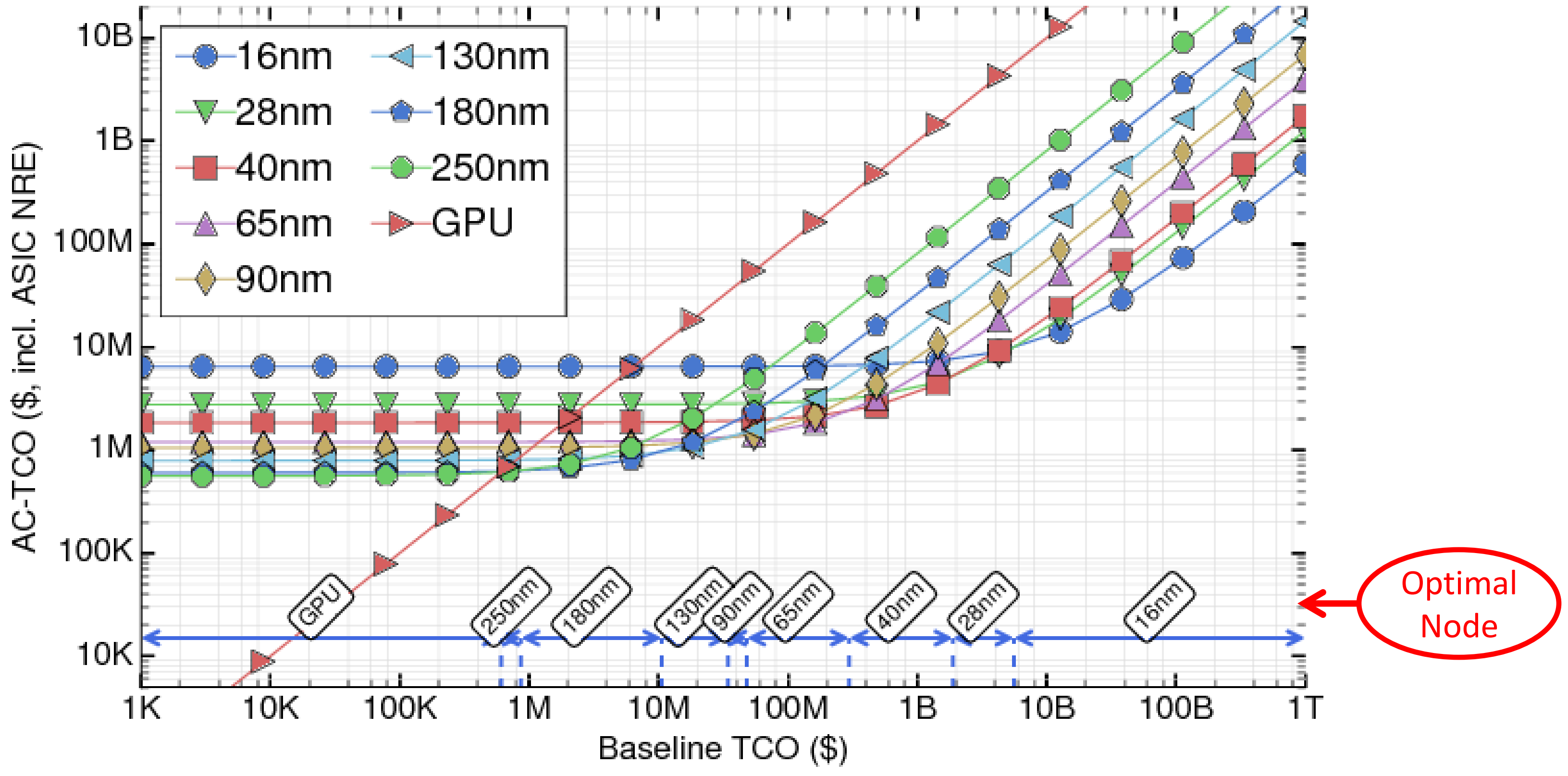
Two-for-Two rule in Practice

- Post-40nm has dramatic increase in NRE vs. Marginal Benefit



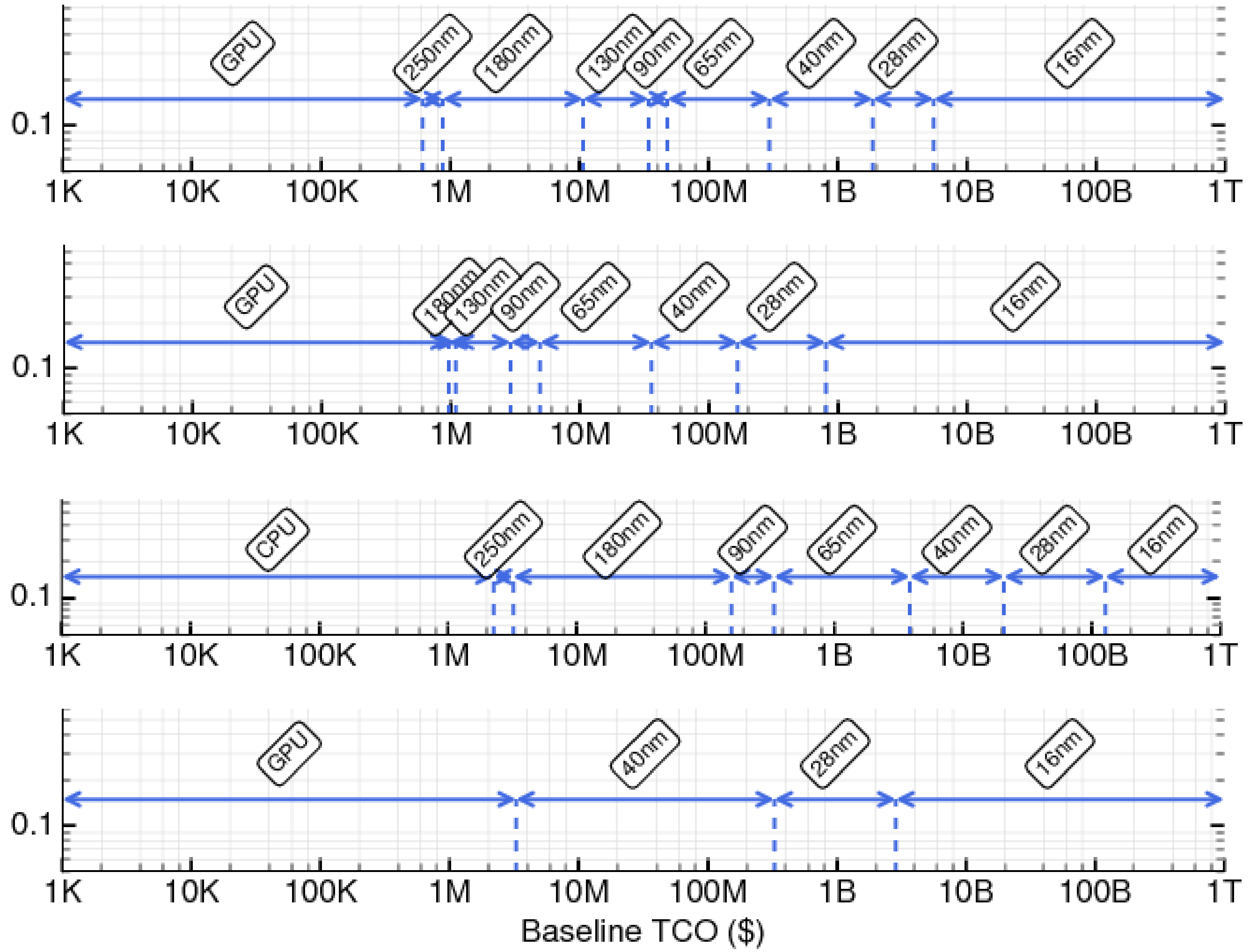
Picking the Optimal Node

- Bitcoin example:



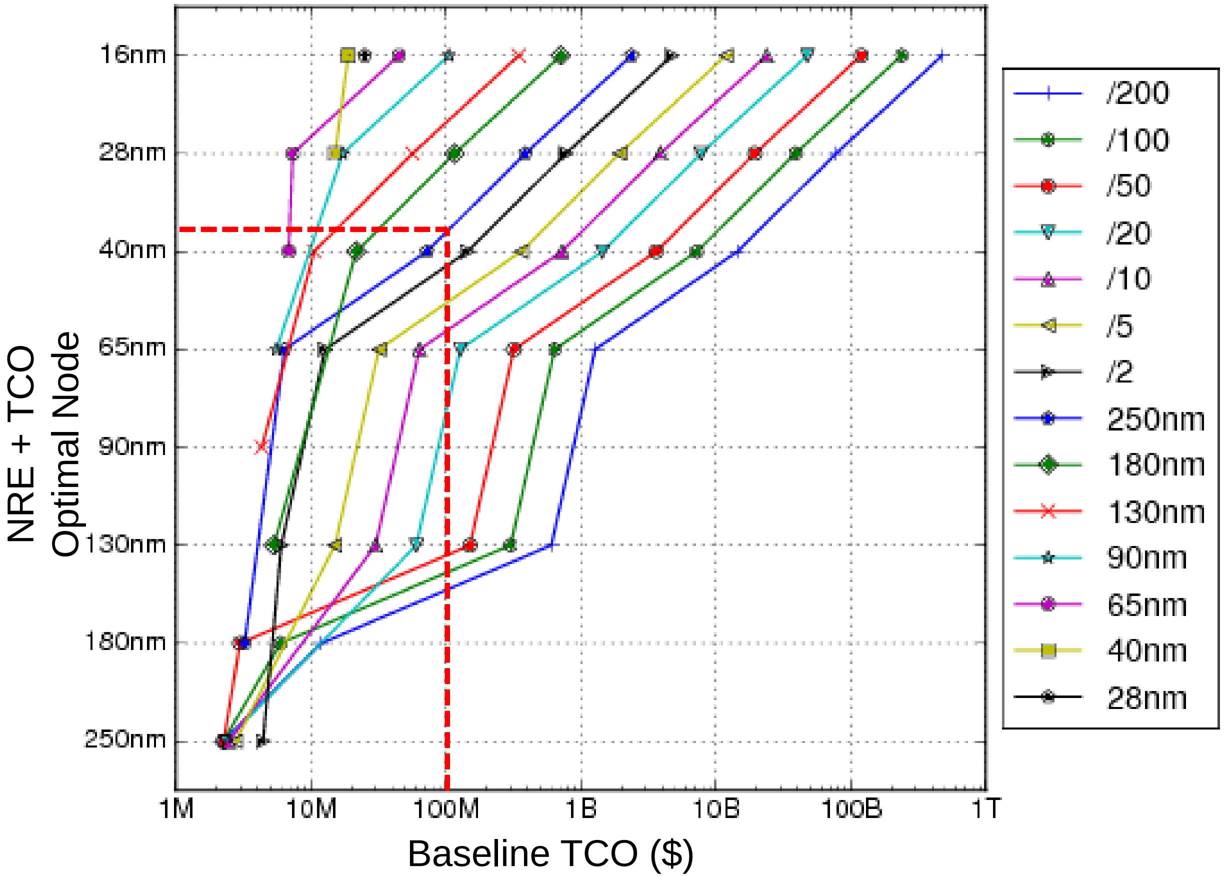
Picking the Optimal Node (cont'd)

- Bitcoin
- Litecoin
- Video Transcode
- Deep Learning



Apps with modest TCO improvement

- Each set of lines represent an app with baseline TCO equal to that ASIC node
- /N means 250 nm reduces TCO by N times
- For time constraints, see the paper how to use it!



Video Transcode Tech Parity Node

Summary

- ASIC Clouds are a promising direction for deploying new kinds of accelerators targeting large, chronic workloads.
- We present a model for computing NRE.
- We present a model for modeling TCO across nodes, and show that old nodes can have optimal NRE+TCO.
- We show an end-to-end methodology for selecting NRE+TCO-optimal ASIC Clouds across technology nodes.

Thank You