Specializing a Planet's Computation: ASIC Clouds

In the last ten years, two parallel phase changes in the computational landscape have emerged. The first change is the bifurcation of computation into two sectors: cloud and mobile; and the second change is the rise of dark silicon and dark silicon aware design techniques such as specialization and near-threshold computation [1].

Recently, researchers and industry have started to examine the conjunction of these two phase changes. GPU-based clouds for distributed neural network accelerators such as Baidu, or FPGA-based clouds deployed by Microsoft for Bing have recently emerged.

At a single node level, we know that ASICs can offer order-magnitude improvements in energy-efficiency and cost-performance over CPU, GPU, and FPGA, by specializing silicon for a particular computation. Our research proposes ASIC Clouds [2], which are purpose-built datacenters comprised of large arrays of ASIC accelerators. ASIC Clouds are not ASIC supercomputers that scale up problem sizes for a single tightly-coupled computation; rather, ASIC Clouds target workloads consisting of many independent but similar jobs.

As more and more services are built around the Cloud model, we see the emergence of planet-scale workloads (think Facebook's face recognition of uploaded pictures, or Apple's Siri voice recognition, or the IRS performing tax audits with neural nets) where datacenters are performing the same computation across many users. These scale-out workloads can easily leverage racks of ASIC servers containing arrays of chips that in turn connect arrays of replicated compute accelerators (RCAs) on an on-chip network. The large scale of these workloads creates the economical justification to pay the non-recurring engineering (NRE) costs of ASIC development and deployment. As a workload grows, the ASIC Cloud can be scaled in the datacenter by adding more ASIC servers, unlike accelerators in say a mobile phone population[3], where the accelerator-to-compute ratio is fixed at tapeout.

Our research examines ASIC Clouds in the context of four key applications that show great potential for ASIC Clouds, including YouTube-style video transcoding, Bitcoin and Litecoin mining, and Deep Learning. ASICs achieve large reductions in silicon area and energy consumption versus CPUs, GPUs, and FPGAs. We specialize the ASIC server to maximize efficiency, employing optimized ASICs, a customized printed circuit board (PCB), custom-designed cooling systems and specialized power delivery systems, and tailored DRAM and I/O subsystems. ASIC voltages are customized in order to tweak energy efficiency and minimize total cost of ownership (TCO). The datacenter itself can also be specialized, optimizing rack-level and datacenter-level thermals and power delivery to exploit the knowledge of the computation. We developed tools that consider all aspects of ASIC Cloud design in a bottom-up

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

way, and methodologies that reveal how the designers of these novel systems can optimize TCO in real-world ASIC Clouds. Finally, we proposed a new rule that explains when it makes sense to design and deploy an ASIC Cloud, considering NRE.

ASIC Cloud Architecture

At the heart of any ASIC Cloud is an energy-efficient, high-performance, specialized *replicated compute accelerator, or RCA,* that is multiplied up by having multiple copies per ASICs, multiple ASICs per server, multiple servers per rack, and multiple racks per datacenter as shown Fig. 1. Work requests from outside the datacenter will be distributed across these RCAs in a scale-out fashion. All system components can be customized for the application to minimize TCO.

ASIC Cloud servers are based on eight enclosed lanes using a duct that has shown to have better cooling performance compared to conventional or staggered layout. Moreover, the baseline server contains a number of DC/DC voltage converters which serve to step down voltage from the power supply unit (PSU) 12 V to a 0.4-1.5 V range.

Scale-out workloads are scheduled to ASICs by means of a control processor, based on a FPGA or CPU. This control unit dispatch computation tasks from outside the server to every ASIC in the on-PCB network, using customized on-PCB multidrop or point-to-point interconnection network, which depending on the required bandwidth can range from serial SPI to parallel LVCMOS or LVDS source synchronous links to complex high speed serial links like HyperTransport or QPI.

Inside every ASIC there is a network of RCAs controlled by a control-plane-unit that process and schedules incoming off-chip packets into the on-chip-network. Moreover, clock generators based on PLL or DLL are added to meet timing requirements for the target application and thermal sensors for monitoring heat. Additionally, we consider the ASIC power-grid as an important design element, since it is tuned for low voltage drop and high current demands of all RCAs. Finally, heatsinks are placed according to the package used by the ASIC. For example, flip-chip packages have heatsinks on top, meanwhile QFN packages have heatsinks on the backside of the PCB.

The sidebar, “Designing an ASIC Cloud”, shows our automated methodology for designing a complete ASIC Cloud system.

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

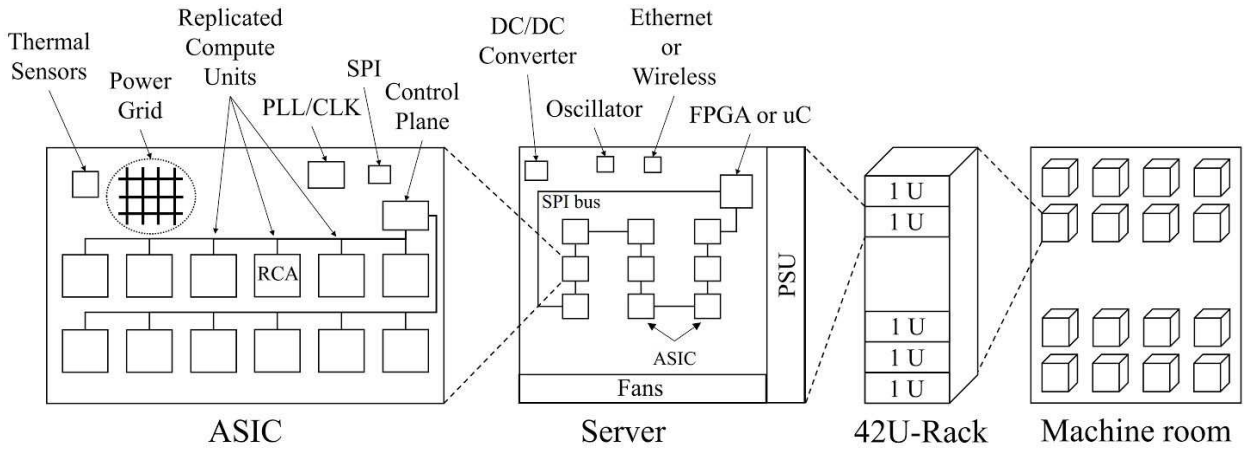
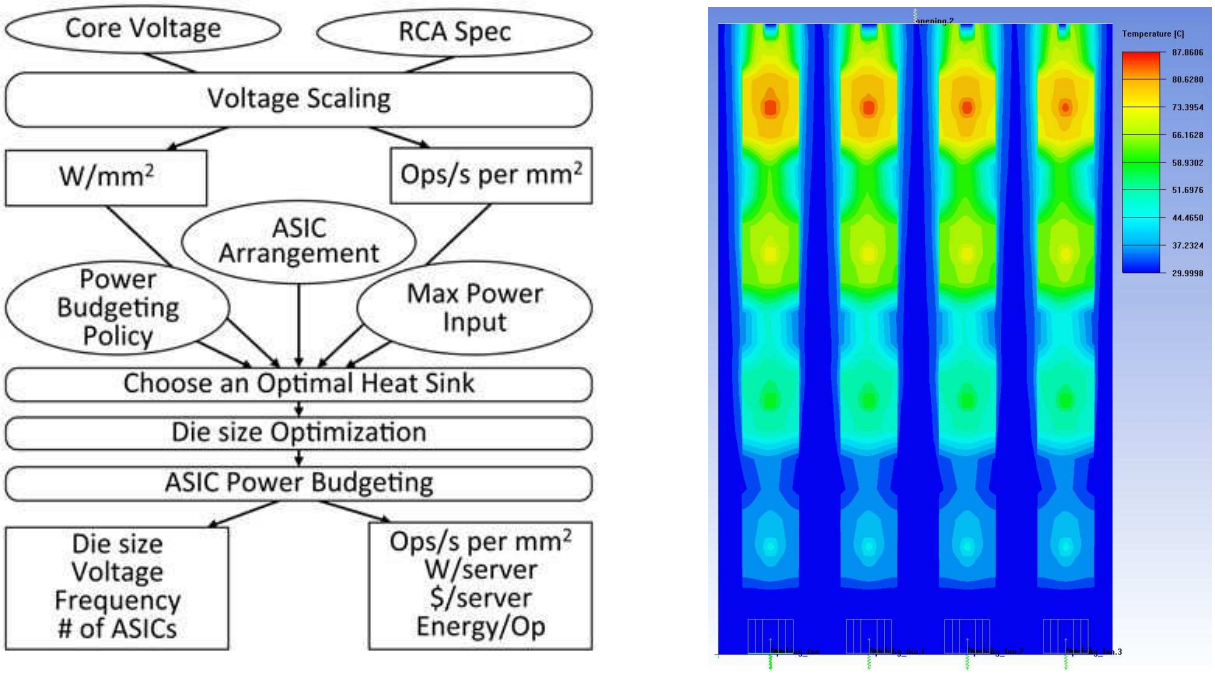


Figure 1. **High-level abstract architecture of an ASIC Cloud.**

BOX 1 : Evaluating an ASIC Server Configuration

Our ASIC Server configuration evaluator, shown in Figure 2a, starts with a Verilog implementation of an accelerator, or a detailed evaluation of the accelerator's properties from the research literature. In the design of an ASIC Server, we must decide how many chips should be placed on the PCB, and how large, in mm^2 of silicon, each chip should be. The size of each chip determines how many RCAs will be on each chip. In each duct-enclosed lane of ASIC chips, each chip receives around the same amount of airflow from the intake fans, but the most downstream chip receives hottest air, which includes the waste heat from the other chips. Therefore, the thermally bottlenecking ASIC is the one in the back, shown in our detailed Computational Fluid Dynamics (CFD) simulations shown in Fig 2b. Our simulations show that breaking a fixed heat source into smaller ones with the same total heat output improves the mixing of warm and cold area, resulting in lower temperatures. Using thermal optimization techniques, we established fundamental connection between an RCA's properties, the number RCA's placed in an ASIC, and how many ASICs go on a PCB in a server. Given these properties, our heat sink solver determines the optimal heat sink configuration. Results are validate with the CFD simulator. In the sidebar entitled "Design Space Evaluation" we show how we apply this evaluation flow across the design space in order to determine TCO and Pareto optimal points that trade off *$ per op/s* and *W per op/s*.

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE



(a) (b)

Figure 2. **ASIC Server Evaluation Flow.** (a) The server cost, per server hash rate, and energy efficiency are evaluated using RCA properties, and a flow that optimizes server heatsinks, die size, voltage and power density. (b) Thermal verification of an ASIC Cloud server using CFD tools to validate the flow results. The farthest ASIC from the fan has the highest temperature and is the bottleneck for power per ASIC at a fixed voltage and energy efficiency.

Application case study

To explore ASIC Clouds across a range of accelerator properties, we examined four applications: Bitcoin mining, Litecoin mining, Video Transcoding, and Deep Learning, that span a diverse range of properties, as shown in Fig. 4.

Perhaps the most critical of these applications is Bitcoin mining. Our inspiration for ASIC Clouds came from our intensive study of Bitcoin mining clouds [4], which are one of the first known instances of a real life ASIC Cloud. Fig. 3 shows the massive scale out of the Bitcoin mining workload, which is now operating at the performance of 3.2 billion GPUs. Bitcoin clouds have undergone a rapid ramp from CPU to GPU to FPGA to the most advanced ASIC technology available today. Bitcoin is a very logic intensive design which has high power density and no need for SRAM or external DRAM.

Litecoin is another popular cryptocurrency mining system that has been deployed into clouds. Unlike Bitcoin, it is an SRAM-intensive application which has low power density.

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

Video Transcoding, which converts from one video format to another, currently takes almost 30 high-end Xeon servers to do in real-time. Since every cell phone can easily be a video source, as well as every IoT device, it has the potential to be an unimaginably large planet-scale computation. Video Transcoding is an external memory-intensive application that needs DRAMs next to each ASIC, and also high off-PCB bandwidth.

Finally, Deep Learning is extremely compute intensive and is likely to be used by every human on the planet. Deep Learning is often latency sensitive so our Deep Learning neural net accelerator has a tight low-latency SLA.

For our Bitcoin and Litecoin studies, we developed the RCA and got the required parameters such as gate count from placed and routed designs in UMC 28nm using Synopsys IC compiler, and analysis tools (e.g. PrimeTime). For Deep Learning and Video Transcoding, we extract properties from accelerators designed in the research literature.

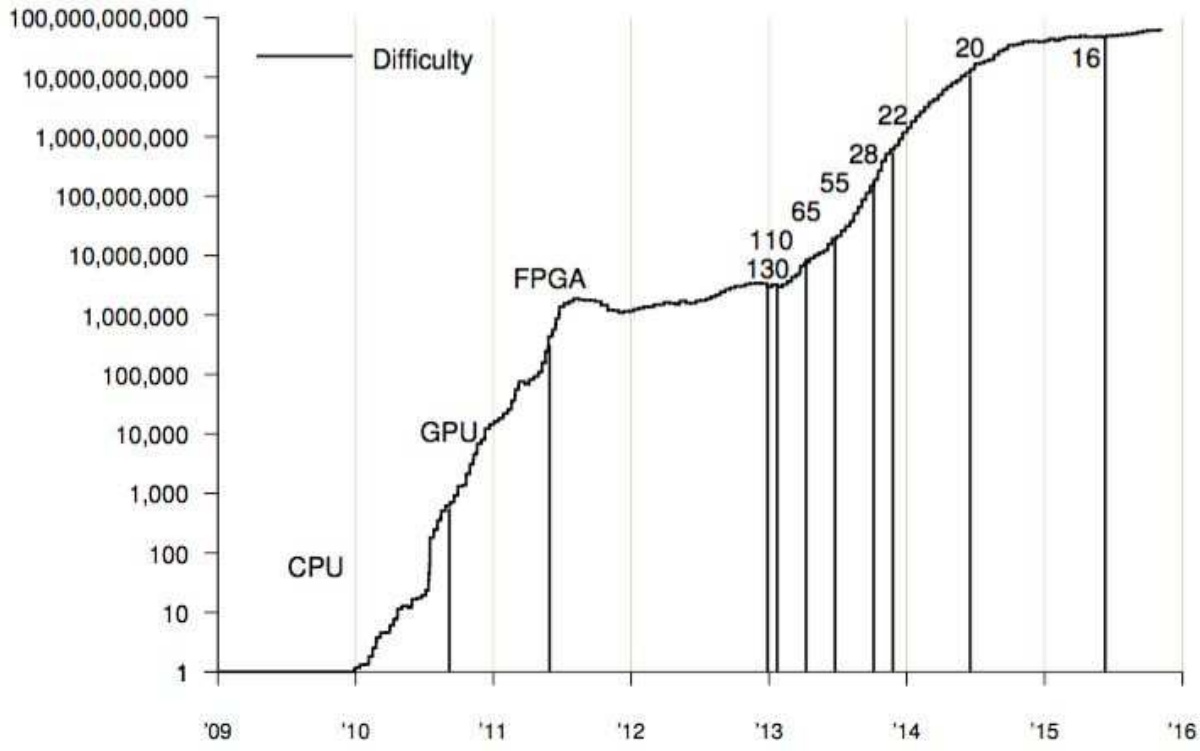


Figure 3. **Evolution of Specialization, Bitcoin cryptocurrency mining clouds.** Numbers are ASIC nodes, in nm, which annotate the first date of release of a miner on that technology. Difficulty is the ratio of the total Bitcoin hash throughput of the world, relative to the initial mining network throughput, which was 7.15 MH/s. In the six-year period preceding Nov 2015, the throughput has increased by a factor of 50 billion times, corresponding to a world hash rate of approximately 575 million GH/s.

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

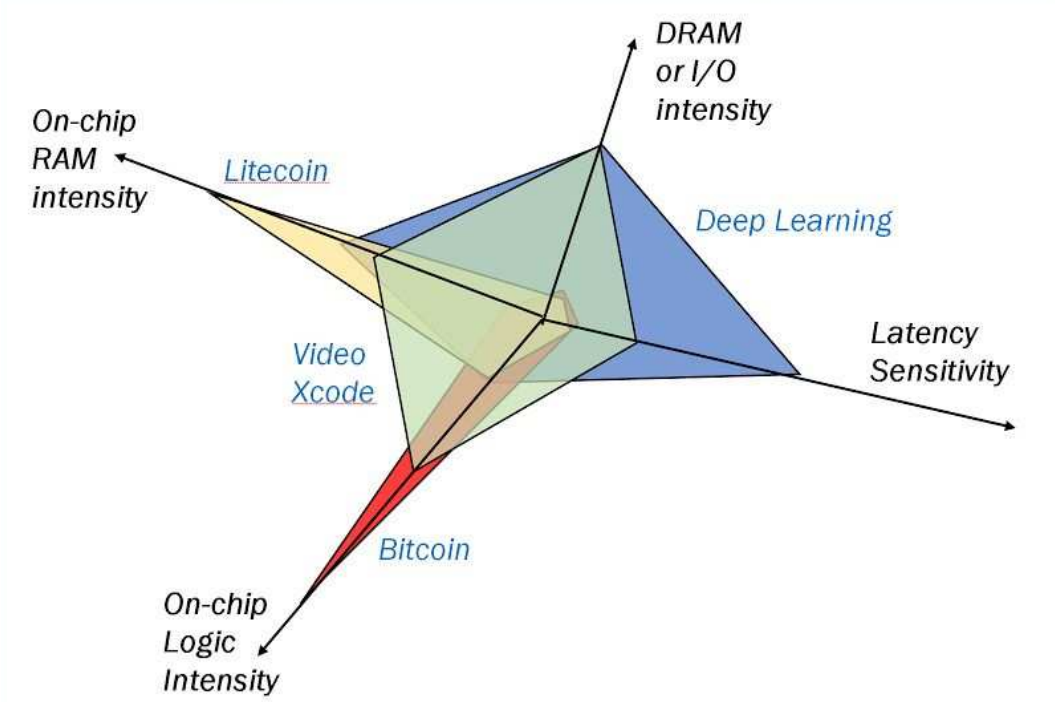


Figure 4. **Accelerator properties.** We explore applications with diverse requirements.

BOX 2: Design space exploration

After having all thermal constraints in place, we optimized ASIC server design targeting two conventional key metrics, namely cost per op/s and power per op/s, and then apply TCO analysis. TCO analysis incorporates the datacenter-level constraints including the cost of power delivery inside the datacenter, land, depreciation, interest, and the cost of energy itself. With these tools, we can correctly weight these two metrics and find the overall optimal point (TCO-optimal) for the ASIC Cloud.

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

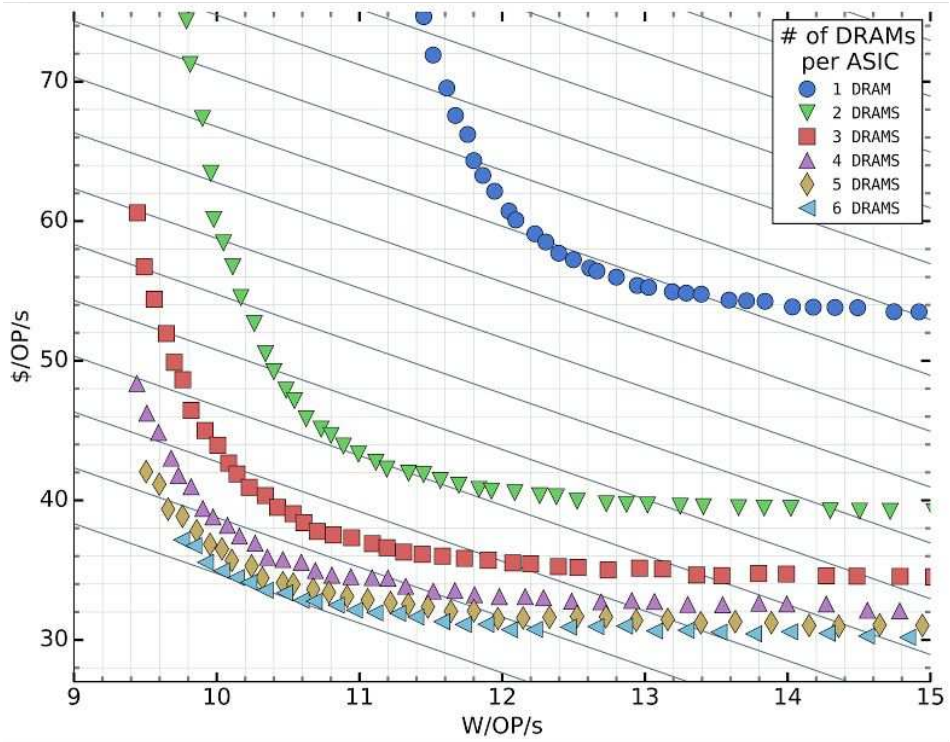


Figure 5. **Pareto curve example for Video Transcode.** Exploring different number of DRAMs per ASIC and logic voltage for optimal TCO per performance point. Voltage increases from left to right. Diagonal lines show equal TCO per performance values and the closer to the origin the lower the TCO per performance. This plot is for 5 ASICs per lane.

Design space exploration is application dependant and there are frequently additional constraints. For example for video transcode application, we model the PCB real estate occupied by these DRAMs, which are placed on either side of the ASIC they connect to, perpendicular to airflow. As the number of DRAMs increases, the number of ASICs placed in a lane decreases for space reasons. We model the more expensive PCBs required by DRAM, with more layers and better signal/power integrity. We employ two 10-GigE ports as the off-PCB interface for network-intensive clouds, and model the area and power of the memory controllers.

Our ASIC Cloud infrastructure explores a comprehensive design space, including DRAMs per ASIC, logic voltage, area per ASIC, and number of chips. DRAM cost and power overhead are significant, and so the Pareto-optima Video Transcoder designs ensure DRAM bandwidth is saturated, linked chip performance to DRAM count. As voltage and frequency is lowered, area increases to meet the performance requirement. Fig. 5 shows the Video Transcode Pareto curve for 5 ASICs per lane and different number of DRAMs per ASIC. The tool is composed of two tiers. The top tier uses brute force to explore all of the possible configurations in order to find the energy-optimal, cost optimal and TCO optimal points are chosen based on the Pareto results. The leaf tier consists of a variety of “expert solvers” that compute optimal properties of the server components; for example, CFD simulations for heat sinks, DC-DC converter

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

allocation, circuit area/delay/voltage/energy estimators, and DRAM property simulation. In many cases these solvers export their data as large tables of memoized numbers for every component.

Results

Details of optimal server configurations for energy-optimal, TCO-optimal and cost-optimal designs for each of the applications are shown in Fig. 6.

	Energy optimal	TCO optimal	Cost optimal
ASICs per server	120	72	24
Logic Voltage (V)	0.400	0.459	0.594
Clock Freq. (MHz)	71	149	435
Die Area (mm2)	599	540	240
GH/s/server	7,292	8,223	3,451
W/server	2,645	3,736	2,513
$/server	12,454	8,176	2,458
W/GH/s	0.363	0.454	0.728
$/GH/s	1.708	0.994	0.712
TCO/GH/s	3.344	2.912	3.686

(a) Bitcoin

	Energy optimal	TCO optimal	Cost optimal
ASICs per server	120	120	72
Logic Voltage (V)	0.459	0.656	0.866
Clock Freq. (MHz)	152	576	823
Die Area (mm2)	600	540	420
MH/s/server	405	1,384	916
W/server	783	3,662	3,766
$/server	10,971	11,156	6,050
W/MH/s	1.934	2.645	4.113
$/MH/s	27.09	8.059	6.607
TCO/MH/s	37.87	19.49	23.70

(b) Litecoin

	Energy optimal	TCO optimal	Cost optimal
DRAMs per ASIC	3	6	9
ASICs per Server	64	40	32
Logic Voltage (V)	0.538	0.754	1.339
Clock Freq. (MHz)	183	429	600
Die Area (mm2)	564	498	543
Kfps/server	126	158	189
W/server	1,146	1,633	3,101
$/server	7,289	5,300	5,591
W/Kfps	9.073	10.34	16.37
$/Kfps	57.68	33.56	29.52
TCO/Kfps	100.3	78.46	97.91

(c) Video Transcode

	Energy optimal	TCO optimal	Cost optimal
Chip type	4x2	2x2	2x1
ASICs per server	32	64	96
Logic Voltage (V)	0.900	0.900	0.900
Clock Freq. (MHz)	606	606	606
TOps/s/server	470	470	353
W/server	3,278	3,493	2,971
$/server	7,809	6,228	4,146
W/TOps/s	6.975	7.431	8.416
$/TOps/s	16.62	13.25	11.74
TCO/TOps/s	46.22	44.28	46.51

(d) Deep learning

Figure 6. **ASIC Cloud Optimization Results for 4 applications.** Each table presents energy-optimal, TCO-optimal and cost optimal server properties. Energy optimal server uses lower voltage to increase the energy efficiency. Cost optimal servers use higher voltage to

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

increase silicon efficiency. TCO-optimal has a voltage between these two and balances energy versus silicon cost.

For example, for Video Transcode, the cost-optimal server packs the maximum number of DRAMs per lane, 36, maximizing performance. However, increasing the number of DRAMs per ASIC requires higher logic voltage (1.34V) and corresponding frequencies to attain performance within the max die area constraint, resulting in less energy efficient designs. Hence, the energy-optimal design has fewer DRAMs per ASIC and per lane (24), while gaining back some performance by increasing ASICs per lane which is possible due to lower power density at 0.54V. The TCO-optimal design increases DRAMs per lane, 30, to improve performance, but is still close to the optimal energy efficiency at 0.75V, resulting in a die size and frequency between the other two optimal points.

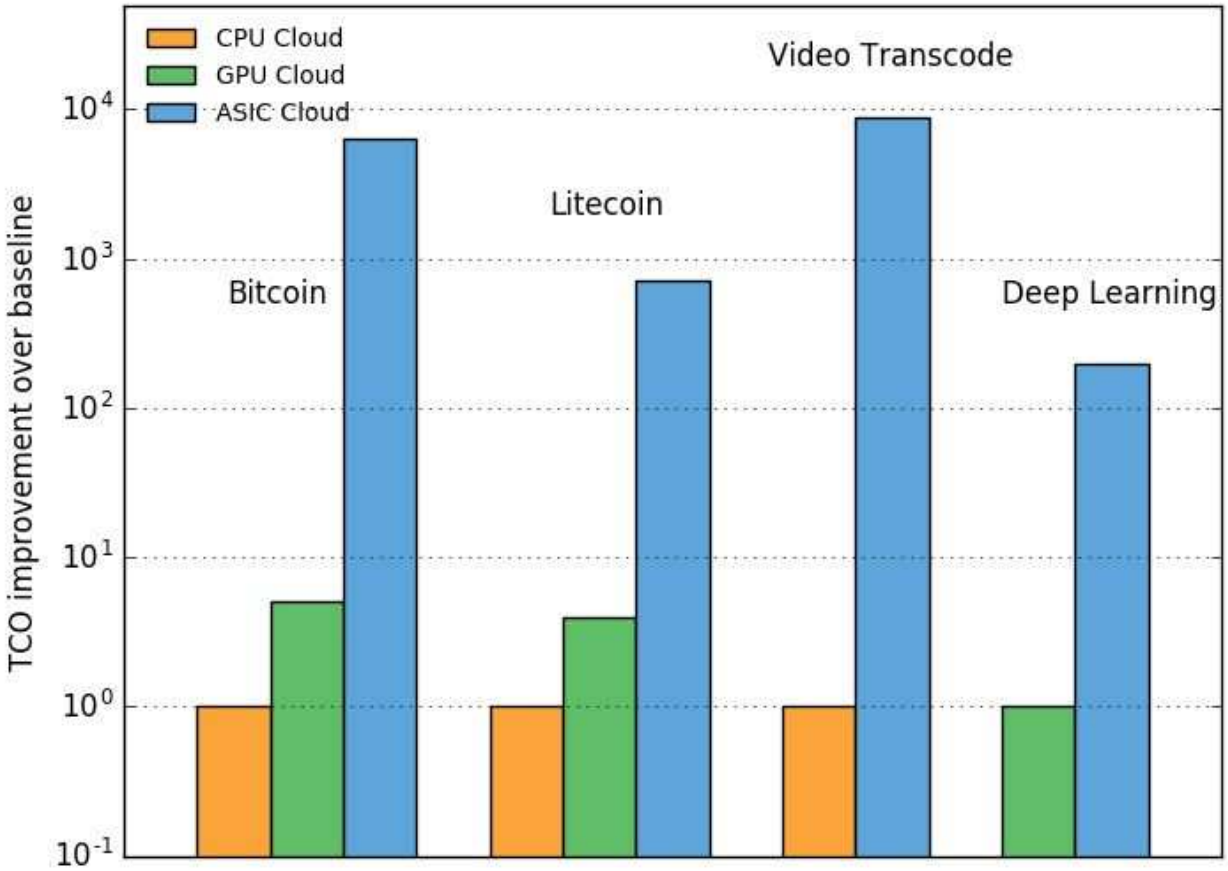


Figure 7. **CPU Cloud vs. GPU Cloud vs. ASIC Cloud Deathmatch.** ASIC Servers greatly outperform the best non-ASIC alternative in terms of TCO per op/s.

In Fig. 7, we compare the performance of CPU Clouds versus GPU Clouds versus ASIC Clouds for the four applications that we presented. ASIC Clouds outperform CPU Cloud TCO per op/s

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

by 6,270x; 704x; and 8,695x for Bitcoin, Litecoin, and Video Transcode respectively. ASIC Clouds outperform GPU Cloud TCO per op/s by 1057x, 155x, and 199x, for Bitcoin, Litecoin, and Deep Learning, respectively.

Feasibility of ASIC clouds: The *two-for-two-rule*

When does it make sense to design and deploy an ASIC Cloud? The key barrier is the cost of developing the ASIC Server, which includes both the mask costs (about $ 1.5M for the 28 nm node we consider here), and the ASIC design costs, which collectively comprise the non recurring engineering expense (NRE). To understand this tradeoff we proposed the ***two-for-two rule***. If the cost per year (i.e. the TCO) for running the computation on an existing cloud exceeds the NRE by 2X, and you can get at least a 2X TCO per operation/second improvement, then going ASIC Cloud is likely to save money. Fig. 8 shows a wider range of breakeven points. Essentially, as the TCO exceeds the NRE by more and more, the required speedup to break even declines. As a result, almost any accelerator proposed in the literature, no matter how modest the speedup, is a candidate for ASIC Cloud, depending on the scale of the computation. Our research makes the key contribution of noting that in deployment of ASIC Clouds, NRE and scale can be more determinative than the absolute speedup of the accelerator. The main barrier for ASIC Clouds is to reign in NRE costs so they are appropriate for the scale of the computation. In many research accelerators, TCO improvements are extreme (such as in Fig 8), but authors often unnecessarily target expensive, latest generation process nodes because they are more cutting edge. This tendency raises the NRE exponentially, reducing economic feasibility. A better strategy is to target the older nodes that still attain sufficient TCO improvements.

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

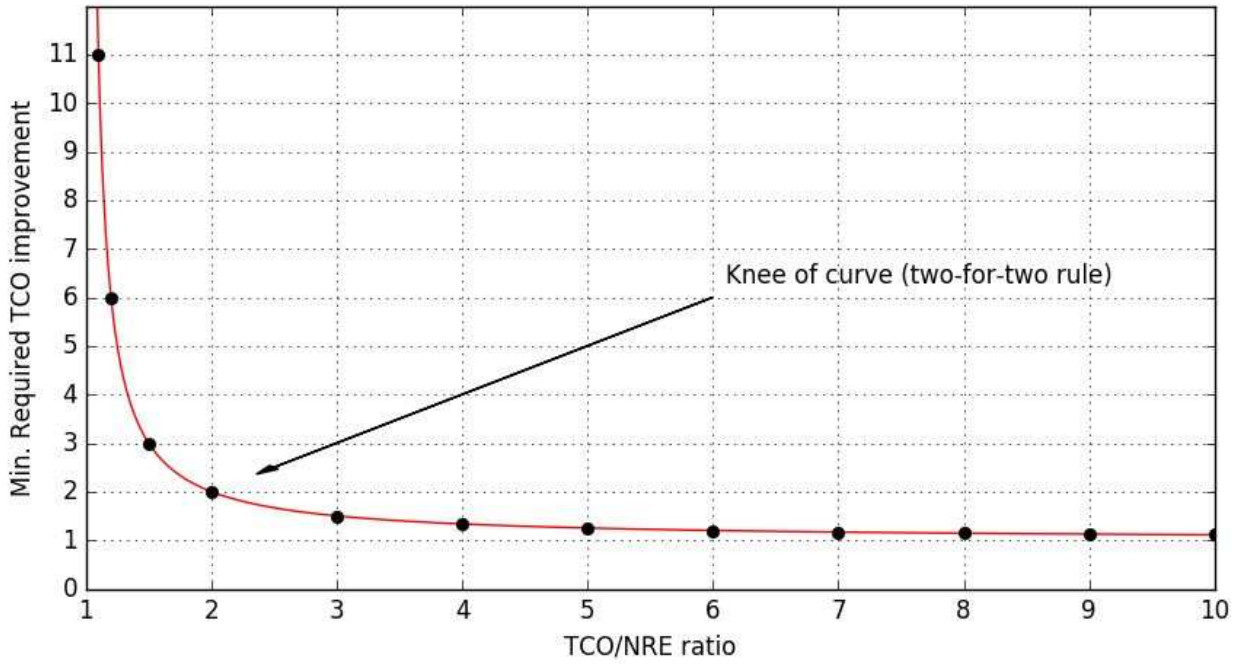


Figure 8. **Two-for-two rule: Moderate speed-up with low NRE beats high speed-up at high NRE.** The points are break even points for ASIC Clouds.

Conclusion

Our research generalizes primordial Bitcoin ASIC Clouds into an architectural template that can apply across a range of planet-scale applications.
We demonstrated methodologies that can be used to design TCO-optimal clouds, answering long standing questions even in contemporary Bitcoin ASIC Clouds. Our work analyses the impact of NRE and scale on deployment of ASIC Clouds, tying it to the TCO-improvement and in turn the energy and cost efficiency of the cloud.

Our work advances research practice by showing how to examine accelerators at a systems level instead of at the level of a single chip. We evaluate ASIC Cloud chip design, server design, and finally datacenter design in a cross-layer system-oriented way. This joint knowledge and control over datacenter and hardware design allows for ASIC Cloud designers to select the optimal design that optimizes energy and cost proportionally. We developed the tools and revealed how the designers of these novel systems can optimize the TCO in real-world ASIC Clouds.

We developed a rule of thumb for when it makes sense to go ASIC Cloud, the *two-for-two rule*. The main barrier for ASIC Clouds is to reign in NRE costs so they are appropriate for the scale of the computation. In many research accelerators, TCO improvements are extreme, but

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE

authors also target expensive, latest generation process nodes because they are more cutting edge. But this habit raises the NRE exponentially, reducing economic feasibility. Our most recent work [5] suggests that a better strategy is to lower NRE cost by targeting older nodes that still have sufficient TCO per op/s benefit.

Looking to the future, our work suggests that both Cloud providers and silicon foundries would benefit by investing in technologies that reduce the NRE of ASIC design, including open source IP such as RISC-V, in new labor-saving development methodologies for hardware and also in open source backend CAD tools. With time, mask costs fall by themselves, but currently older nodes such as 65 nm and 40 nm may provide suitable TCO per op/s reduction, with half the mask cost and only a small difference in performance and energy efficiency from 28 nm. Foundries should take interest in ASIC Cloud's low-voltage scale out design patterns because they lead to greater silicon wafer consumption than CPUs within fixed environmental energy limits.

With the coming explosive growth of planet-scale computation, we must work to contain the exponentially growing environmental impact of datacenters across the world. ASIC Clouds promise to help address this problem. By specializing the datacenter, they can do greater amounts of computation under environmentally determined energy limits. The future is planet-scale, and specialized ASICs will be everywhere.

References

[1] M. B. Taylor. A Landscape of the Dark Silicon Design Regime. IEEE Micro, 2013.

[2] I. Magaki et al. ASIC Clouds: Specializing the Datacenter. ISCA, 2016.

[3] N. Goulding-Hotta et al. The GreenDroid Mobile Application Processor: An Architecture for Silicon's Dark Future. IEEE Micro, 2011.

[4] M. B. Taylor. Bitcoin and the age of bespoke silicon. CASES, 2013.

[5] M. Khazraee et al. Moonwalk: NRE Optimization in ASIC Clouds or, accelerators will use old silicon. ASPLOS, 2017.

Digital Object Identifier 10.1109/MM.2017.265085658 0272-1732/$26.00 2017 IEEE