Specializing a Planet's Computation: ASIC Clouds

ASIC Clouds, a natural evolution to CPU- and GPU-based clouds, are purpose-built datacenters filled with ASIC accelerators. ASIC Clouds may seem improbable due to high non-recurring engineering (NRE) costs and ASIC inflexibility, but large-scale Bitcoin ASIC Clouds already exist. This article distills lessons from these primordial ASIC Clouds and proposes new planet-scale YouTube-style video-transcoding and Deep Learning ASIC clouds, showing superior total cost of ownership. ASIC Cloud NRE and economics are also examined.

Moein Khazraee
Luis Vega Gutierrez
Ikuo Magaki
Michael Bedford Taylor
University of California, San Diego

In the past 10 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. 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. Baidu has developed GPU-based clouds for distributed neural network accelerators, and Microsoft has deployed clouds based on field-programmable gate arrays (FPGAs) for Bing.

At a single-node level, we know that application-specific integrated circuits (ASICs) can offer order-of-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 comprising large arrays of ASIC accelerators. ASIC Clouds are not ASIC supercomputers that scale up problem sizes for a single tightly coupled computation; rather, they target workloads comprising many independent but similar jobs.

As more and more services are built around the Cloud model, we see the emergence of planet-scale workloads in which datacenters are performing the same computation across many users. For example, consider Facebook's face recognition of uploaded pictures, or Apple's Siri voice recognition, or the Internal Revenue Service performing tax audits with neural nets. Such 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 economic justification to pay the non-recurring

Published by the IEEE Computer Society

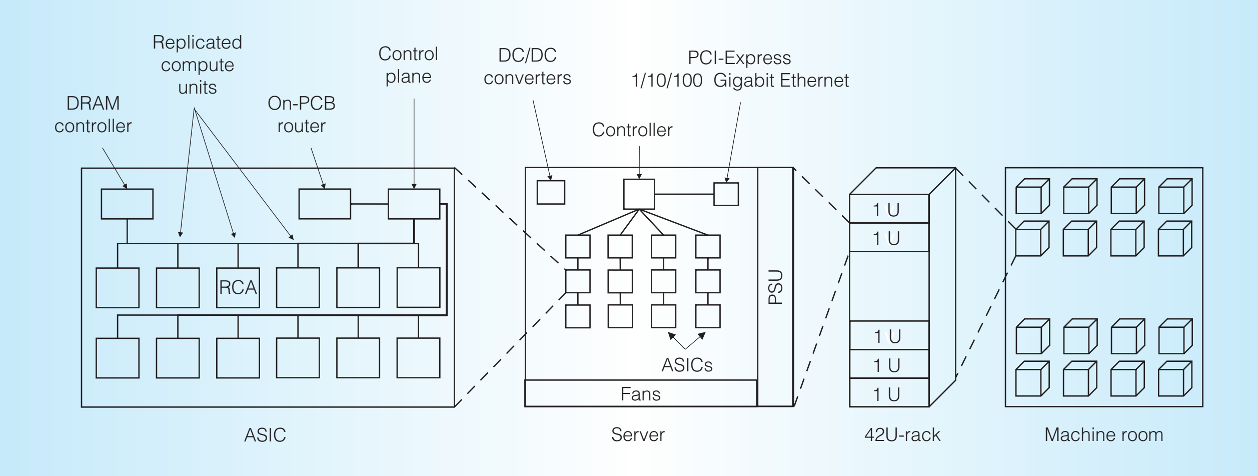

Figure 1. High-level abstract architecture of an ASIC Cloud. Specialized replicated compute accelerators (RCAs) are multiplied up by having multiple copies per application-specific integrated circuit (ASIC), multiple ASICs per server, multiple servers per rack, and multiple racks per datacenter. Server controller can be a field-programmable gate array (FPGA), microcontroller, or a Xeon processor. The power delivery and cooling system are customized based on ASIC needs. If required, there would be DRAMs on the printed circuit board (PCB) as well. (PSU: power supply unit.)

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] in which the accelerator/processor mix is fixed at tape out.

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 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 way, and methodologies that reveal how the designers of these novel systems can optimize TCO in real-world ASIC Clouds. Finally, we propose 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 RCA that is multiplied up by having multiple copies per ASIC, multiple ASICs per server, multiple servers per rack, and multiple racks per datacenter (see Figure 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.

Each ASIC interconnects its RCAs using a customized on-chip network. The ASIC's control plane unit also connects to this network and schedules incoming work from the ASIC's off-chip router onto the RCAs. Next, the packaged ASICs are arranged in lanes on a customized PCB and connected to a controller that bridges to the off-PCB interface (1 to 100 Gigabit Ethernet, Remote Direct Memory Access, and PCI Express). In some cases, DRAMs can connect directly to the ASICs. The controller can be implemented by an FPGA, a microcontroller, or a Xeon processor. It schedules remote procedure calls (RPCs) that come from the off-PCB interface on to the ASICs.

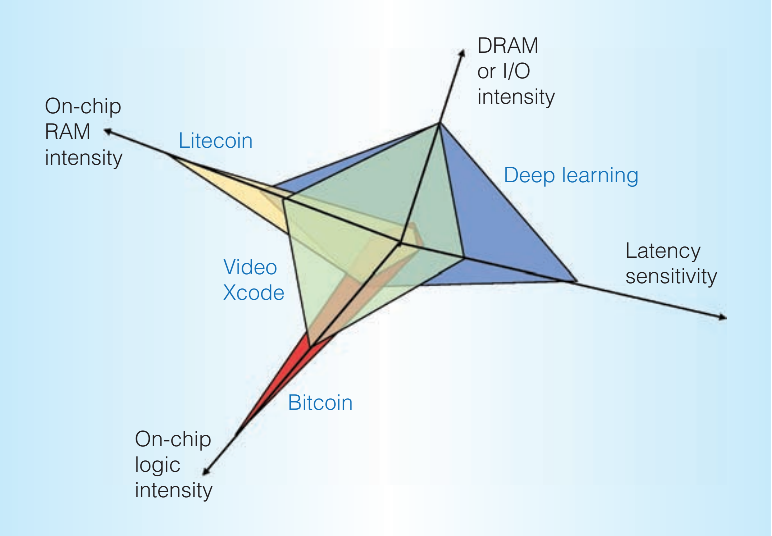

Figure 2. Accelerator properties. We explore applications with diverse requirements.

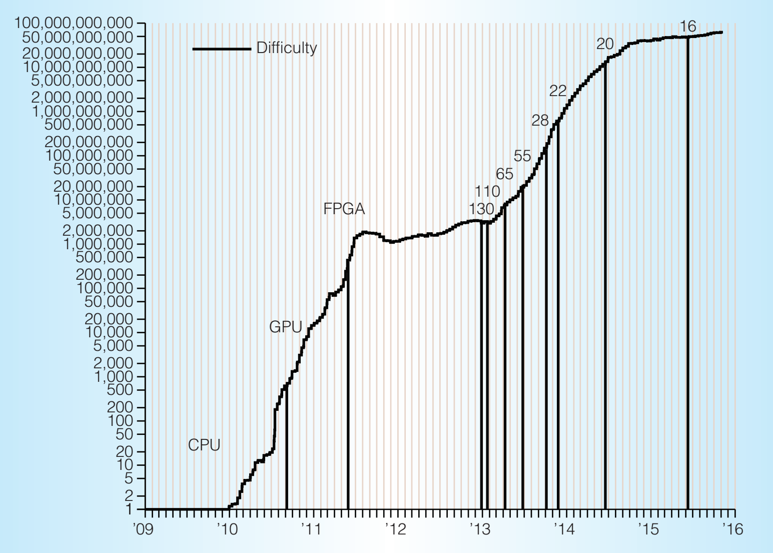

Figure 3. Evolution of specialization: Bitcoin cryptocurrency mining clouds. Numbers are ASIC nodes, in nanometers, which annotate the first date of release of a miner on that technology. Difficulty is the ratio of the total Bitcoin hash throughput of the world, relative to the initial mining network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

Depending on the application, it can implement the non-acceleratable part of the workload or perform UDP/TCP-IP offload.

Each lane is enclosed by a duct and has a dedicated fan blowing air through it across the ASIC heatsinks. Our simulations indicate using ducts results in better cooling performance compared to conventional or staggered layout. The PCB, fans, and power supply are enclosed in a 1U server, which is then assembled into racks in a datacenter. Based on ASIC needs, the power supply unit (PSU) and DC/DC converters are customized for each server.

The "Evaluating an ASIC Server Configuration" sidebar shows our automated methodology for designing a complete ASIC Cloud system.

Application Case Study

To explore ASIC Clouds across a range of accelerator properties, we examined four applications that span a diverse range of properties—namely, Bitcoin mining, Litecoin mining, video transcoding, and deep learning (see Figure 2).

Perhaps the most mature 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. Figure 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 logic-intensive design that has high power density and no need for static RAM (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 with low power density.

Video transcoding, which converts from one video format to another, currently takes almost 30 high-end Xeon servers to do in real time. Because every cell phone and Internet of Things device can easily be a video source, 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. It also requires high off-PCB bandwidth.

Finally, deep learning is extremely computationally intensive and is likely to be used by every human on the planet. It is often latency sensitive, so our Deep Learning neural

Evaluating an ASIC Server Configuration

Our ASIC Cloud server configuration evaluator, shown in Figure A1, 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 printed circuit board (PCB) and how large, in mm^2 of silicon, each chip should be. The size of each chip determines how many replicated compute accelerators (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 the 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 in Figure A2. 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 areas, resulting in lower temperatures. Using thermal optimization techniques, we established a fundamental connection between an RCA's properties, the number of RCAs 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 validated with the CFD simulator. In the "Design Space Evaluation" sidebar, we show how we apply this evaluation flow across the design space to determine TCO and Pareto-optimal points that trade off cost per operation per second (ops/s) and watts per ops/s.

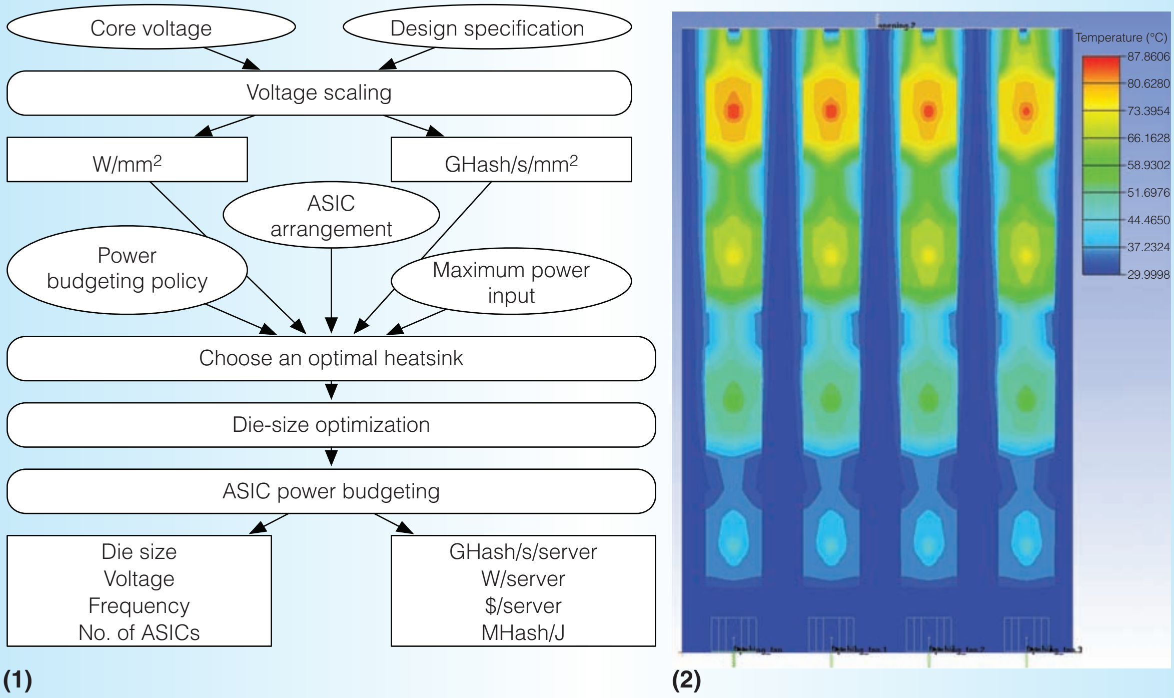

Figure A. ASIC server evaluation flow. (1) The server cost, per server hash rate, and the energy efficiency are evaluated using replicated compute accelerator (RCA) properties and a flow that optimizes server heatsinks, die size, voltage, and power density. (2) Thermal verification of an ASIC Cloud server using Computational Fluid Dynamics 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.

net accelerator has a tight low-latency service-level agreement.

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 28 nm using Synopsys IC compiler and analysis tools (such as PrimeTime). For deep learning and video transcoding, we extracted properties from accelerators in the research literature.

Design Space Exploration

After all thermal constraints were in place, we optimized ASIC server design targeting two conventional key metrics—namely, cost per ops/s and power per ops/s—and then applied 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.

Design-space exploration is application dependent, and there are frequently additional constraints. For example, for the video transcoding 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-Gigabit Ethernet ports as the off-PCB interface for network-intensive clouds, and we 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-optimal video transcoding designs ensure DRAM bandwidth is saturated, and link chip performance to DRAM count. As voltage and frequency are lowered, area increases to meet the performance requirement. Figure B shows the video transcoding Pareto curve for five ASICs per lane and different numbers of DRAMs per ASIC. The tool comprises two tiers. The top tier uses brute force to explore all possible configurations to find the energy-optimal, cost-optimal, and TCO-optimal points based on the Pareto results. The leaf tier comprises various expert solvers that compute the optimal properties of the server components—for example, CFD simulations for heat sinks, DC-DC converter 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.

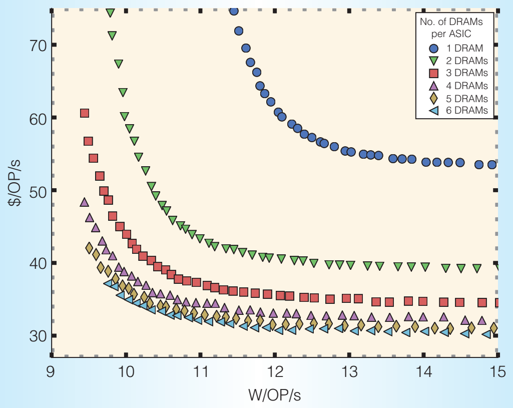

Figure B. Pareto curve example for video transcoding. Exploring different numbers of DRAMs per ASIC and logic voltage for optimal TCO per performance point. Voltage increases from left to right. Diagonal lines show equal TCO per performance values; the closer to the origin, the lower the TCO per performance. This plot is for five ASICs per lane.

Results

Table 1 gives details of optimal server configurations for energy-, TCO-, and cost-optimal designs for each application. The “Design Space Exploration” sidebar explains how these optimal configurations are determined.

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

Figure 4 compares the performance of CPU Clouds, GPU Clouds, and ASIC Clouds for the four applications that we presented. ASIC Clouds outperform CPU Clouds’ TCO per operations per second (ops/s) by 6,270, 704, and 8,695 times for Bitcoin, Litecoin, and video transcoding, respectively. ASIC Clouds outperform GPU Clouds’ TCO per ops/s by 1,057, 155, and 199 times for Bitcoin, Litecoin, and deep learning, respectively.

ASIC Cloud Feasibility: The Two-for-Two Rule

When does it make sense to design and deploy an ASIC Cloud? The key barrier is

Table 1. ASIC Cloud optimization results for four applications: (a) Bitcoin, (b) Litecoin, (c) video transcoding, and (d) deep learning.

Property	Energy optimal server	TCO optimal server	Cost optimal server
ASICs per server	120	72	24
Logic voltage (V)	0.400	0.459	0.594
Clock frequency (MHz)	71	149	435
Die area (mm^2)	599	540	240
GH per second (GH/s) per server	7,292	8,223	3,451
W per server	2,645	3,736	2,513
Cost ($) per server	12,454	8,176	2,458
W per GH/s	0.363	0.454	0.728
Cost ($) per GH/s	1.708	0.994	0.712
Total cost of ownership (TCO) per GH/s	3.344	2.912	3.686
(a)			
ASICs per server	120	120	72
Logic voltage (V)	0.459	0.656	0.866
Clock frequency (MHz)	152	576	823
Die area (mm^2)	600	540	420
MH/s per server	405	1,384	916
W per server	783	3,662	3,766
$ per server	10,971	11,156	6,050
W per MH/s	1.934	2.645	4.113
$ per MH/s	27.09	8.059	6.607
TCO per MH/s	37.87	19.49	23.70
(b)			
DRAMs per ASIC	3	6	9
ASICs per server	64	40	32
Logic voltage (V)	0.538	0.754	1.339
Clock frequency (MHz)	183	429	600
Die area (mm^2)	564	498	543
Kilo frames per second (Kfps) per server	126	158	189
W per server	1,146	1,633	3,101
$ per server	7,289	5,300	5,591
W per Kfps	9.073	10.34	16.37
$ per Kfps	57.68	33.56	29.52
TCO per Kfps	100.3	78.46	97.91
(c)			
Chip type	4 × 2	2 × 2	2 × 1
ASICs per server	32	64	96
Logic voltage (V)	0.900	0.900	0.900
Clock frequency (MHz)	606	606	606
Tera-operations per second (Tops/s) per server	470	470	353
W per server	3,278	3,493	2,971
$ per server	7,809	6,228	4,146
W per Tops/s per server	6.975	7.431	8.416
$ per Tops/s per server	16.62	13.25	11.74
TCO per Tops/s per server	46.22	44.28	46.51
(d)			

*Energy-optimal server uses lower voltage to increase the energy efficiency. Cost-optimal server uses higher voltage to increase silicon efficiency. TCO-optimal server has a voltage between these two and balances energy versus silicon cost.

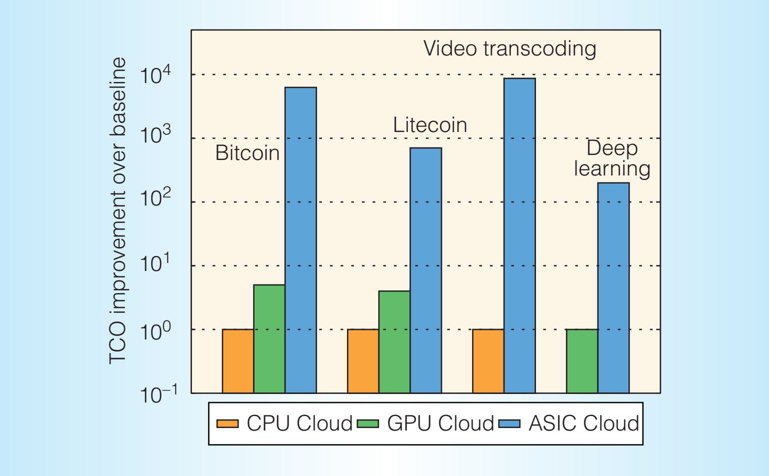

Figure 4. CPU Cloud versus GPU Cloud versus ASIC Cloud death match. ASIC servers greatly outperform the best non-ASIC alternative in terms of TCO per operations per second (ops/s).

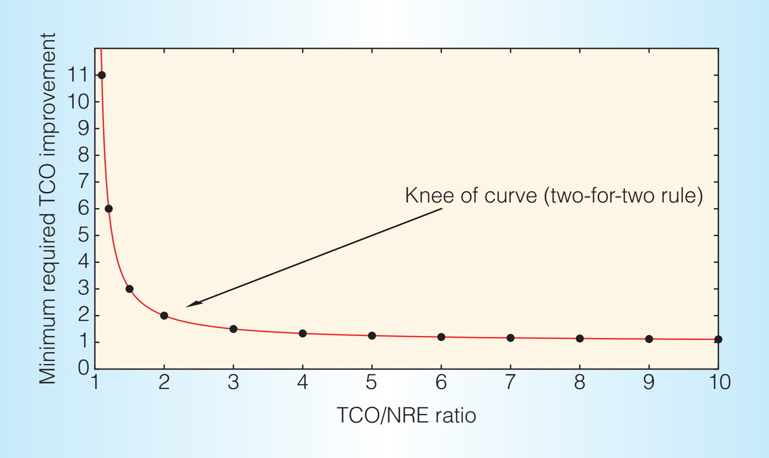

Figure 5. The two-for-two rule. Moderate speedup with low non-recurring engineering (NRE) beats high speedup at high NRE. The points are break-even points for ASIC Clouds.

the cost of developing the ASIC server, which includes both the mask costs (about $1.5 million for the 28-nm node we consider here), and the ASIC design costs, which collectively comprise the NRE expense. To understand this tradeoff, we proposed the two-for-two rule. If the cost per year (that is, the TCO) for running the computation on an existing cloud exceeds the NRE by two times, and you can get at least a two-times TCO improvement per ops/s, then building an ASIC Cloud is likely to save money.

Figure 5 shows a wider range of break-even 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 the 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 Figure 4), 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. Our most recent work suggests that a better strategy is to lower NRE cost by targeting older nodes that still have sufficient TCO per ops/s benefit.[5]

Our research generalizes primordial Bitcoin ASIC Clouds into an architectural template that can apply across a range of planet-scale applications. 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 to optimize TCO. 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 in open source back-end CAD tools. With time, mask costs fall by themselves, but older nodes such as 65 nm and 40 nm may provide suitable TCO per ops/s reduction, with one-third to half the mask cost and only a small difference in performance and energy efficiency from 28 nm. This is a major shift from the conventional wisdom in architecture research, which often chooses the best process even though it exponentially increases NRE. Foundries also 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. MICRO

Acknowledgments

This work was partially supported by NSF awards 1228992, 1563767, and 1565446, and by STARnet's Center for Future Architectures Research, a SRC program sponsored by MARCO and DARPA.

References

1. M.B. Taylor, "A Landscape of the Dark Silicon Design Regime," *IEEE Micro*, vol. 33, no. 5, 2013, pp. 8–19.
2. I. Magaki et al., "ASIC Clouds: Specializing the Datacenter," *Proc. 43rd Int'l Symp. Computer Architecture*, 2016, pp. 178–190.
3. N. Goulding-Hotta et al., "The GreenDroid Mobile Application Processor: An Architecture for Silicon's Dark Future," *IEEE Micro*, vol. 31, no. 2, 2011, pp. 86–95.
4. M.B. Taylor, "Bitcoin and the Age of Bespoke Silicon," *Proc. Int'l Conf. Compilers, Architectures and Synthesis for Embedded Systems*, 2013, article 16.
5. M. Khazraee et al., "Moonwalk: NRE Optimization in ASIC Clouds," *Proc. 22nd Int'l Conf. Architectural Support for Programming Languages and Operating Systems*, 2017, pp. 511–526.

Moein Khazraee is a PhD candidate in the Department of Computer Science and Engineering at the University of California, San Diego. His research interests include ASIC Clouds, NRE, and specialization. Khazraee received an MS in computer science from the University of California, San Diego. Contact him at mkhazraee@ucsd.edu.

Luis Vega Gutierrez is a staff research associate in the Department of Computer Science and Engineering at the University of California, San Diego. His research interests include ASIC Clouds, low-cost ASIC design, and systems. Vega received an MSc in electrical and computer engineering from the University of Kaiserslautern, Germany. Contact him at lvgutierrez@ucsd.edu.

Ikuo Magaki is an engineer at Apple. He performed the work for this article as a Toshiba visiting scholar in the Department of Computer Science and Engineering at the University of California, San Diego. His research interests include ASIC design and ASIC Clouds. Magaki received an MSc in computer science from Keio University, Japan. Contact him at ikuo.magaki@icloud.com.

Michael Bedford Taylor advises his PhD students at various well-known west coast universities. He performed the work for this article while at the University of California, San Diego. His research interests include tiled multicore architecture, dark silicon, HLS accelerators for mobile, Bitcoin mining hardware, and ASIC Clouds. Taylor received a PhD in electrical engineering and computer science from the Massachusetts Institute of Technology. Contact him at profmbt@uw.edu.