Overclocking in Immersion-Cooled Datacenters

Pulkit A. Misra, *Microsoft Research, Redmond, WA, 98052, USA*
Ioannis Manousakis, *Microsoft Azure, Redmond, WA, 98052, USA*
Esha Choukse, *Microsoft Research, Redmond, WA, 98052, USA*
Majid Jalili, *The University of Texas at Austin, Austin, TX, 78712, USA*
Íñigo Goiri, *Microsoft Research, Redmond, WA, 98052, USA*
Ashish Raniwala and Brijesh Warrier, *Microsoft Azure, Redmond, WA, 98052, USA*
Husam Alissa and Bharath Ramakrishnan, *Microsoft CO+I, Redmond, WA, 98052, USA*
Phillip Tuma, *3M Company, Saint Paul, MN, 55144, USA*
Christian Belady, *Microsoft CO+I, Redmond, WA, 98052, USA*
Marcus Fontoura, *Microsoft Azure, Redmond, WA, 98052, USA*
Ricardo Bianchini, *Microsoft Research, Redmond, WA, 98052, USA*

Large cloud providers are starting to leverage liquid cooling for an increasing number of workloads. Liquid cooling enables providers to overclock server components, but they must tradeoff the potential increase in performance with higher power draw and reliability implications. We argue that two-phase immersion cooling is the most promising technology and, in that context, explore overclocking, its uses, and implications.

Cloud providers typically use air-based cooling in datacenters as it is currently easy to install and operate. However, air cooling has a few critical downsides. First, its heat dissipation efficiency is low, and it thus requires large heat sinks and fans that increase space and costs. Second, operating at higher component junction temperatures results in higher leakage power, which in turn negatively impacts energy efficiency. Most importantly, increasing transistor counts and the end of Dennard scaling will soon result in chips with thermal design power (TDP) that will make air cooling essentially inviable.[11]

For these reasons, providers have started to explore liquid cooling (e.g., cold plates, immersion cooling) for their most power-hungry workloads.[1] These technologies keep chip temperatures at a lower and narrower range than air cooling, reducing leakage power, eliminating the need for fans, and reducing datacenter power usage effectiveness (PUE). For example, Google cools its tensor processing units (TPUs) with cold plates.[2] Alibaba introduced single-phase immersion cooling tanks in their datacenters and showed that it reduces the total power consumption by 36% and achieves a PUE of 1.07.[12]

LIQUID COOLING ENABLES PROVIDERS TO OPERATE SERVER COMPONENTS BEYOND THE NORMAL FREQUENCY RANGE (I.E., OVERCLOCK THEM) FOR LONGER PERIODS OF TIME THAN EVER POSSIBLE BEFORE.

Liquid cooling enables providers to operate server components beyond the normal frequency range (i.e., overclock them) for longer periods of time than

0272-1732 © 2022 IEEE
Digital Object Identifier 10.1109/MM.2022.3163107
Date of publication 5 April 2022; date of current version 30 June 2022.

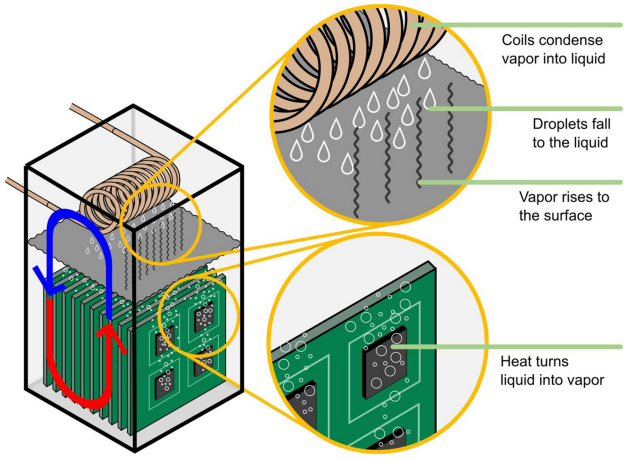

FIGURE 1. Two-phase immersion cooling. Hardware is submerged into a dielectric liquid that changes phase (boils). Vapor rises to the top where it rejects heat and condenses back to liquid form. This process requires no additional energy. (Source: Allied Control Limited; used with permission.)

ever possible before. This overclocking capability opens many new directions to enhance system performance and customer experience at scale.

However, overclocking does not come for free, as it increases power consumption and can impact component reliability. Worse, overclocking might not even improve performance for some workloads. For example, overclocking the central processing unit (CPU) running a memory-bound workload will not result in much performance gain. For these reasons, providers must carefully manage the benefits, risks, and costs of overclocking.

In this article, a summary of our ISCA'21 paper,[5] we explore immersion and the ability to overclock while managing the risks. We start by comparing cooling technologies and argue that two-phase immersion cooling (2PIC) is the most promising of them. We then describe our 2PIC tank prototypes. Given the benefits of 2PIC, we discuss the various aspects of overclocking, including power, component lifetime, and total cost of ownership (TCO) implications. Next, we propose use cases for cloud providers to take advantage of overclocking. Finally, we conclude by discussing the possible long-term impact of our work.

Immersion Cooling in Datacenters

Cloud providers have recently started to employ liquid cooling. Their initial efforts typically focused on placing cold plates on power-hungry components, where fluid flows through the plates and the piping to remove heat. Although efficient, each cold plate needs to be specifically designed for each new component, which increases engineering complexity and time to market. Moreover, air cooling is still required for the other components on a server.

An alternative to cold plates is immersion cooling, where entire servers are submerged in a tank and the heat is dissipated by direct contact with a dielectric liquid. The heat removal can happen in a single- or two-phase manner. In single-phase immersion (1PIC), the tank liquid absorbs the heat and circulates using pumps, whereas in 2PIC a phase-change process from liquid to vapor (via boiling) carries the heat away, as illustrated in Figure 1.

Table 1 compares the power and thermal efficiency of the different air and liquid cooling technologies. The overheads with air cooling (three top rows) will increase with higher server power but remain stable with liquid cooling (three bottom rows). 2PIC achieves the lowest PUE due to its more efficient heat transfer. Furthermore, unlike cold plates, 2PIC does not require repeated engineering for each new server hardware generation. Finally,

TABLE 1. Comparison of the main datacenter cooling technologies.[5]

Cooling Technology	Average PUE	Peak PUE	Server fan overhead	Max server cooling
Chillers	1.70	2.00	5%	700 W
Water-side	1.19	1.25	6%	700 W
Direct evaporative	1.12	1.20	6%	700 W
CPU cold plates	1.08	1.13	3%	2 kW
1PIC	1.05	1.07	0%	2 kW
2PIC	1.02	1.03	0%	>4 kW

FIGURE 2. Large 2PIC tank prototype.

2PIC can support a higher thermal flux than 1PIC. Thus, we argue that 2PIC is the most promising technology.

Our 2PIC tank prototypes: To study immersion and aggressive component overclocking, we build three prototypes: 1) two small tanks able to host two servers each and 2) a large tank hosting 36 servers (see Figure 2). Each tank contains different combinations of liquids (and boiling points) and server equipment, so we can experiment broadly with 2PIC and overclocking. Our earlier paper[5] presents results from many of these experiments. Most importantly, we started using the large tank in a production environment[9] and will expand this program significantly over the next few years.

Overclocking in Immersion Cooling

Figure 3 shows the allowable frequency ranges for server-class processors today. Most times, the processors operate within the guaranteed range between the minimum and the base frequency. In air-cooled datacenters, they can opportunistically operate at turbo frequency when thermal and power budgets allow. For example, Intel offers Turbo Boost v2.0,[4] which opportunistically increases core frequency depending on the thermal headroom, number of active cores, and type of instructions executed. In contrast, 2PIC has very high cooling capability and thereby provides greater opportunities for overclocking, irrespective of utilization and without a significant impact on temperatures.

However, overclocking has several side-effects that must be traded off carefully against its potential performance benefits. Next, we discuss the important aspects that need to be considered.

Power consumption: Power is an important datacenter consideration as it increases infrastructure cost. Although overclocking increases power consumption, the power efficiency improvements (lower leakage power from reduced operating temperatures, elimination of server fans, and lower PUE) through 2PIC can significantly offset the increase.

Despite these savings, indiscriminate overclocking may result in hitting limits in the datacenter's power delivery infrastructure and triggering capping mechanisms. These mechanisms rely on CPU frequency reduction and memory bandwidth throttling for controlling power and therefore might offset any performance gains from overclocking.

Component lifetime: Increasing the operating frequency (and consequently voltage) may reduce the lifetime of electronics. However, immersion lowers the operating temperature and can compensate for the lifetime impact of overclocking. Our evaluation[5] with a 5-nm composite processor model from a major fabrication company shows that an overclocked processor in 2PIC has the same lifetime of a nonoverclocked air-cooled processor.

Computational stability: Overclocking may induce bit flips due to aggressive circuit timing and sudden voltage drops. Bit flips can cause applications to crash or produce erroneous results (silent errors). Fortunately, processors already implement error correction mechanisms.[3] Furthermore, safe overclocking limits can be established in coordination with vendors to avoid instability and relevant counters can be monitored during runtime for safety.

Environmental impact: Overclocking may increase overall energy consumption and be an indirect source of

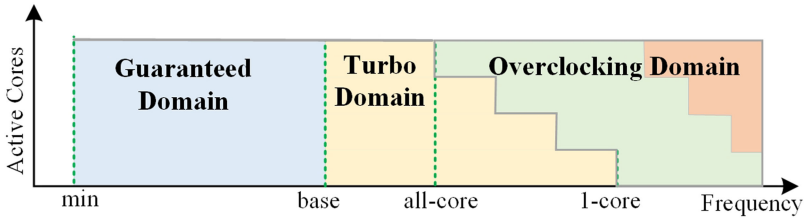

FIGURE 3. Operating domains: guaranteed (blue), turbo (yellow), and overclocking (green). The nonoperating domain is marked with red.

TABLE 2. Normalized test load profiles for the microsoft teams conference-serving workload.

Load profile	# Audio-only calls	# Audio-video calls			Total users
	4 users/call (Small AO)	4 users/call (Small AV)	50 users/call (Medium AV)	250 users/call (Large AV)	
Low	1x	1x	1x	0	1x
Medium	1x	1x	1x	1x	1.5x
High	2x	2x	2x	0	2x

CO_2 emissions. However, as datacenters will be primarily powered by renewables,[7] this effect will become less relevant. For immersion, we are exploring liquids with very low global warming impact and ensuring that tanks are sealed and have careful vapor management.

Total cost of ownership: 2PIC adds costs for tanks and liquid that are offset by the elimination of fans and reduction in PUE. The lower PUE enables adding more servers to the datacenter, thereby amortizing costs (e.g., construction, operations, energy). Our analysis[5] shows that nonoverclockable 2PIC datacenters are 7% cheaper than an air-cooled baseline. The TCO savings reduce to 4% with overclocking. This is because: 1) the server power infrastructure needs to be upgraded for overclocking and 2) overclocking may increase energy consumption. However, using overclocking to enable greater oversubscription (one of the use cases we propose below) produces TCO savings of 13% compared to the baseline.

Performance: The impact of overclocking depends on the workload's bottleneck resource. Despite the thermal benefits of immersion, providers must carefully use overclocking to provide performance benefits when they are achievable while managing the associated risks and costs.

Overclocking Use Cases in Datacenters

We propose several overclocking use cases to enhance customer experience and/or reduce costs:

1) offer high-performance virtual machines (VMs);
2) improve packing density of VMs on servers (including by using more aggressive oversubscription);
3) reduce capacity buffers;
4) mitigate capacity crises;
5) enhance VM autoscaling services.

Next, we discuss a few of the use cases in more detail.

High-performance VMs: Cloud providers today offer VMs with Turbo support.[8] However, with the ability to overclock, a provider could offer new high-performance VM classes that run at even higher frequencies. Using our tank prototypes and realistic cloud workloads, we quantified the performance versus power tradeoffs of overclocking CPU cores, last-level caches (LLCs), memory, and graphics processing units (GPUs).[5] A key result is that overclocking improves performance between 10% and 25% for the workloads, with the highest (lowest) benefit coming from core (LLC) overclocking. Memory and LLC overclocking often increase average and 99th-percentile power significantly for little or no performance benefit for the workloads we studied.

With a high-performance VM offering, customers can get more performance from each VM and thereby save costs through reducing the number of VMs needed for handling high loads. As a concrete example, we evaluate this benefit using the conference-serving benchmark from Microsoft Teams. The workload has a client-server setup, where clients are the end-users participating in conference calls and the server instances run on VMs in Azure. A server instance can handle multiple conference calls and acts as a relay for the calls' participants. Moreover, each call has multiple participants (ranging from 2 to several hundreds) and can be audio-only or audio-video, with different codecs and video resolutions. The benchmark allows specifying the number of calls hosted by the server VM, the type (audio-only vs audio-video) of each call, and the number of users (clients) per call.

Table 2 shows the normalized load profiles that we use for evaluation. The workload performance is measured by the following indicators on the server VM:

1) CPU utilization;
2) conference processing rate;
3) bitrate—bandwidth consumed by the incoming and outgoing channels to the server;
4) number of overloaded conferences—the call quality is degraded for these conferences; lower values are better for 1 and 4 and higher for 2 and 3.

Finally, the duration of each run (conference time emulation) is ten minutes, and the results are averaged over three runs.

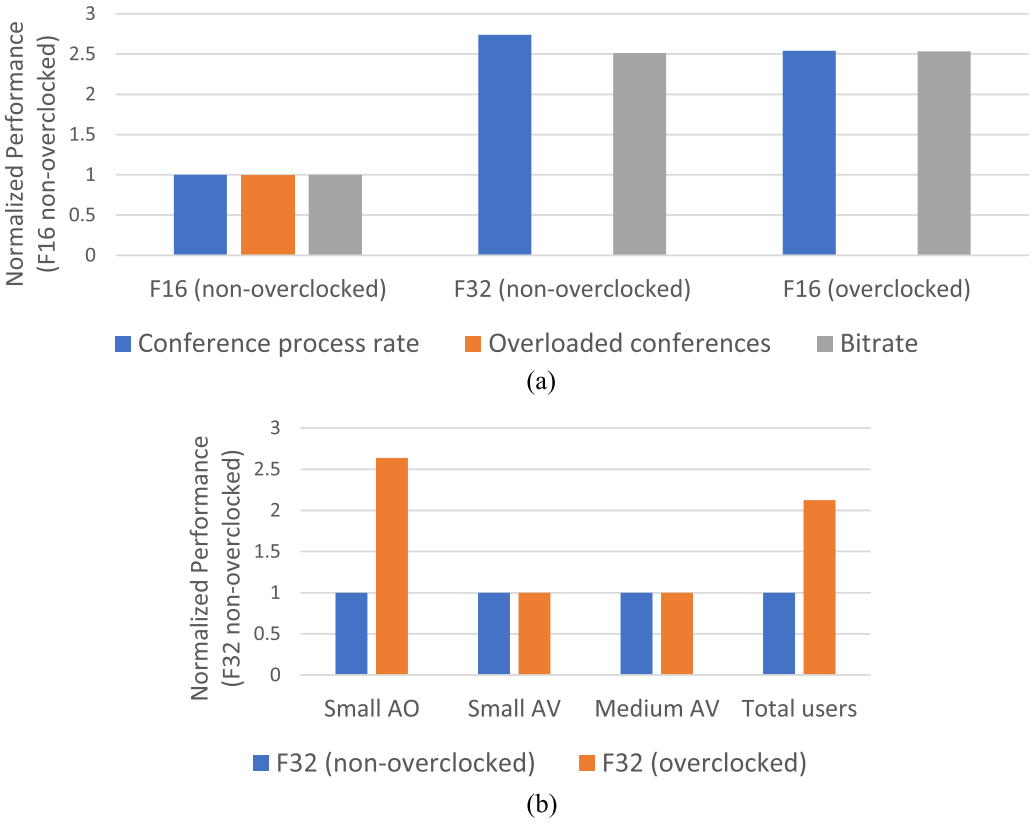

FIGURE 4. Impact of using high-performance VMs by the Microsoft Teams conference-serving workload. (a) Impact of using overclocking with a fixed load profile. (b) Impact of using overclocking on a fixed-size server VM.

Since the workload is CPU-heavy, we use the compute-optimized F-series[8] VM offerings from Azure. For the server VM, we use either an F16 or F32 VM; F16 comes with 16 vCPUs and 32 GB of dynamic random-access memory (DRAM) and F32 is the next larger offering with double the number of vCPUs and memory. We evaluate the benefit of overclocking for the workload by changing the core frequency of the server VM from 2.6 to 3.4 GHz.

We start by evaluating the performance impact of overclocking for all three load profiles when the server instance is running on a F16 VM. Compared to the baseline, the overclocked version provides up to 28% reduction in CPU utilization on average, while offering a 2.5× increase in conference processing rate across all three load profiles. During the evaluation, we observe that the nonoverclocked server VM gets saturated with the high load profile, and this causes 11% of the conferences to be in an overloaded state. Since the customer experience is impacted because of the overloading, the workload owner will need to run such a load mix on a larger F32 VM to not have an impact on call quality. Figure 4(a) compares the performance offered by the nonoverclocked F32 and the overclocked F16 VM for the high load profile; the performance numbers are relative to the nonoverclocked F16 VM. As the results indicate, overclocking provides similar performance while using a smaller-sized (F16 versus F32) VM.

Next, we evaluate the impact of overclocking on the load-bearing ability of a VM. We use an F32 VM for this experiment and evaluate how many conference calls the VM can handle without and with overclocking, while having no conference calls in an overloaded state. We start with the high-load profile and keep increasing the number of four-user audio-only conference calls to find the maximum number of calls that the VM can host without degrading call quality. Figure 4(b) shows the results: overclocking the server VM enables handling 2.3× additional overall conferences and 2.1× total users when compared to the nonoverclocked version.

Dense VM packing: Providers use multidimensional bin packing to place VMs on servers. To protect

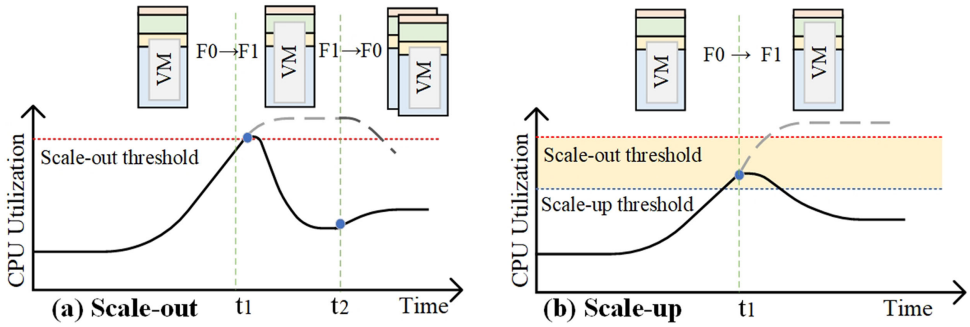

FIGURE 5. Using overclocking to improve autoscaling.

performance, they must pack so that the VMs on a server are unlikely to need the same resources at the same time. However, when this rare scenario occurs, providers can use overclocking to mitigate it. In fact, this mitigation mechanism can even enable an increase in VM packing density (VMs/server ratio) and thereby cut costs by reducing the number of servers required. Even a single percentage point in tighter VM packing equates to hundreds of millions of dollars in savings for large cloud providers like Azure.[10] Our evaluation[5] shows that providers can increase VM packing density by 20% when overclocking is combined with CPU oversubscription. Furthermore, our TCO analysis shows that increasing density by just 10% when coupled with overclocking would reduce the cost per virtual core for Azure by 13% in comparison to today's air-cooled scenario.

Overclocking-enhanced autoscaling: Providers offer services for autoscaling the number of VMs in a deployment according to user-provided rules.[6] For example, users may specify that a new VM should be added (scale-out) if the average CPU utilization over the last 5 min exceeds 50%. However, scaling-out can take tens of seconds to even minutes to deploy new VMs. This overhead imposes a performance penalty on applications, forcing them to scale-out before the extra VMs are needed.

Instead, providers can temporarily overclock the existing VMs (i.e., scale-up) to mitigate the performance impact of scale-out. Furthermore, scaling-up can be used proactively to prevent the need for scaling-out altogether and thereby reduce customer cost. Figure 5(a) and (b) illustrates these two applications of overclocking for autoscaling, respectively. Our evaluation[5] shows that overclocking can improve the tail latency of a latency-sensitive workload by 54% compared to a traditional autoscaling system, while also reducing the number of VMs needed for the workload when overclocking is used proactively to prevent scale-out.

Potential Long-Term Impact

We expect our work to have a strong and lasting impact on both academia and industry for many reasons.

> *RESEARCHERS CAN DESIGN COMPONENT WEAR-OUT COUNTERS TO DYNAMICALLY TRACK THE RELIABILITY BUDGET AND EXPLORE ITS USE FOR TRADING OFF BETWEEN OVERCLOCKING INTENSITY, DURATION, AND COMPONENT LIFETIME.*

Inexorable need for liquid cooling: Hardware trends indicate that many chips will soon have TDPs beyond the capabilities of cost-effective air cooling. Thus, multiple cloud providers are exploring liquid cooling. Our work introduces computer architects to immersion cooling and the opportunity to rethink cloud computing in an environment where thermal constraints are drastically relaxed. The implications are enormous and span datacenter, server, and device design. For example, immersion enables higher server density, which would reduce datacenter footprints and enable more efficient resource disaggregation. In addition, low operating temperatures enable building processors that use faster but leakier transistors. Effective cooling can also ease the path for three-dimensional (3-D) components and layouts.

New research prompted by immersion: From these examples, it should be obvious that eliminating thermal constraints opens multiple research avenues for computer architects in academia and industry: How dense can we make servers? What are the server repair implications of such dense designs? How much faster can

we make resource disaggregation? Can disaggregated memory have on-board NUMA-level latency? What kinds of 3-D server layouts can we conceive? How much faster can we make cores, memory, networking, and storage?

Systematic evaluation of 2PIC for datacenters: Our work compares multiple cooling technologies and shows that 2PIC is a promising option. This comparison guided our decision to deploy 2PIC in production and demonstrate the first production-quality 2PIC tank for a cloud provider. Our work will encourage other providers to consider 2PIC.

Enabling overclocking in cloud platforms: While overclocking is not used by cloud platforms today, there are many potential use cases for it. Our work introduces a framework for exploring the trade-offs (e.g., power, component lifetime and stability, performance) and demonstrates the potential benefits, risks, and costs of overclocking.

Our findings can also guide research on enabling overclocking of various components in the system and managing the risks. For example, today, component reliability is calculated statically by hardware vendors based on certain workload and utilization assumptions. However, platforms host a mix of workloads that exhibit a wide range of utilization patterns. Therefore, moderately utilized components will accumulate lifetime credit. Such components can be overclocked further for additional performance, but the extent and duration of overclocking must be balanced against the impact on lifetime. To this end, researchers can design component wear-out counters to dynamically track the reliability budget and explore its use for trading off between overclocking intensity, duration, and component lifetime. Similarly, they can propose policies for managing power within and across servers when overclocking.

Cloud providers define their service-level agreements more broadly than a single hardware component. Researchers can explore hardware-software codesigns for dynamically overclocking the resources associated with such software abstractions (e.g., VM, container). This includes designing fast and efficient interfaces for communicating the performance requirements and priority of the software abstractions to the hardware, while operating under constraints (e.g., power, peak current). These topics can foster collaborations between industry and academia.

Use cases for overclocking in cloud platforms: Our proposed use cases demonstrate the novel space of overclocking in cloud platforms to lower costs and improve customer experience. Industry can take advantage of the use cases and realize the benefits, whereas researchers can leverage our tradeoff framework to propose new use cases.

Adapting overclocking to the cooling technology: By identifying and quantifying its implications, our work opens up the possibility of adapting overclocking to the cooling in cloud platforms. For example, with air cooling, researchers can explore a combination of over and underclocking (depending on the impact of frequency on workload performance), while operating below component thermal limits.

CONCLUSION

In this article, we explored the use of liquid cooling and component overclocking by public cloud providers. We argued that 2PIC is the most promising technology and built three tanks to demonstrate it. We also proposed use-cases for 2PIC-enabled overclocking and discussed the associated benefits and risks. We conclude that 2PIC and overclocking have enormous potential for future cloud platforms.

REFERENCES

1. 3M, "Two-phase immersion cooling a revolution in data center efficiency," 2015. [Online]. Available: https://multimedia.3m.com/mws/media/1127920O/2-phase-immersion-coolinga-revolution-in-datacenter-efficiency.pdf
2. Google, "Cloud tensor processor unit." Accessed: Jun. 5, 2022. [Online]. Available: https://cloud.google.com/tpu
3. Intel, "New reliability, availability, and serviceability (RAS) features in the intel xeon processor family," 2017. [Online]. Available: https://www.intel.com/content/www/us/en/developer/articles/technical/new-reliabilityavailability-and-serviceability-ras-features-in-the-intel-xeon-processor.html
4. Intel, "Turbo boost technology 2.0." Accessed: Jun. 5, 2022. [Online]. Available: https://www.intel.com/content/www/us/en/architecture-and-technology/turbo-boost/turbo-boosttechnology.html
5. M. Jalili *et al.*, "Cost-efficient overclocking in immersion-cooled datacenters," in *Proc. Int. Symp. Comput. Archit.*, 2021, pp. 623–636.
6. Microsoft, "Azure autoscaling." Accessed: Jun. 5, 2022. [Online]. Available: https://azure.microsoft.com/en-us/features/autoscale
7. Microsoft, "Azure sustainability." Accessed: Jun. 5, 2022. [Online]. Available: https://azure.microsoft.com/en-us/global-infrastructure/sustainability
8. Microsoft, "Fsv2-series VMs." [Online]. Available: https://docs.microsoft.com/en-us/azure/virtual-machines/fsv2-series

9. Microsoft, "To cool datacenter servers, microsoft turns to boiling liquid." [Online]. Available: https://news.microsoft.com/innovation-stories/datacenter-liquidcooling/
10. O. Hadary *et al.*, "Protean: VM allocation service at scale," in *Proc. Symp. Oper. Syst. Des. Implementation*, 2020, pp. 845–861.
11. Y. Sun *et al.*, "Summarizing CPU and GPU design trends with product data," 2019. [Online]. Available: https://arxiv.org/abs/1911.11313
12. Y. Zhong, "A large scale deployment experience using immersion cooling in datacenters," Alibaba Group: Open Compute Project Summit, 2019.

PULKIT A. MISRA is a senior research engineer with Microsoft Research, Redmond, WA, 98052, USA. Misra received a Ph.D. degree in computer science from Duke University, Durham, NC, USA. Contact him at pumisra@microsoft.com.

IOANNIS MANOUSAKIS is a principal software engineer with Microsoft Azure, Redmond, WA, 98052, USA. Manousakis received a Ph.D. degree in computer science from Rutgers University, New Brunswick, NJ, USA. Contact him at iomanous@microsoft.com.

ESHA CHOUKSE is a researcher at Microsoft Research, Redmond, WA, 98052, USA. Choukse received a Ph.D. degree in electrical engineering from the University of Texas at Austin, Austin, TX, USA. Contact her at esha.choukse@microsoft.com.

MAJID JALILI is currently working toward a Ph.D. degree in computer science with the University of Texas at Austin, Austin, TX, 78712, USA. He was an intern at Microsoft Research during this work. Contact him at majid@utexas.edu.

ÍÑIGO GOIRI is a principal research software developer with Microsoft Research, Redmond, WA, 98052, USA. Goiri received a Ph.D. degree in computer science from the Universitat Politecnica de Catalunya, Barcelona, Spain. Contact him at inigog@microsoft.com.

ASHISH RANIWALA is a partner architect with Microsoft Azure, Redmond, WA, 98052, USA. Raniwala received a Ph.D. degree in computer science from Stony Brook University, Stony Brook, NY, USA. Contact him at asraniwala@microsoft.com.

BRIJESH WARRIER is a partner engineering manager with Microsoft Azure, Redmond, WA, 98052, USA. He is the engineering lead for efficiency efforts in Azure Compute. Contact him at Brijesh.Ramachandran@microsoft.com.

HUSAM ALISSA is a principal engineer and technical lead with Microsoft CO+I, Redmond, WA, 98052, USA. Alissa received a Ph.D. degree in mechanical engineering from Binghamton University, Binghamton, NY, USA. Contact him at hualissa@microsoft.com.

BHARATH RAMAKRISHNAN is a mechanical engineer at Microsoft CO+I, Redmond, WA, 98052, USA. Ramakrishnan received a Ph.D. degree in mechanical engineering from Binghamton University, Binghamton, NY, USA. Contact him at Bharath.Ramakrishnan@microsoft.com.

PHILLIP TUMA is an advanced application development Engineer at 3M, Saint Paul, MN, 55144, USA, specializing in heat transfer fluids. Contact him at petuma@mmm.com.

CHRISTIAN BELADY is a vice president and distinguished engineer with Microsoft CO+I, Redmond, WA, 98052, USA, where he leads the Datacenter Advanced Development group. Contact him at Christian.Belady@microsoft.com.

MARCUS FONTOURA is a technical fellow with Microsoft Azure, Redmond, WA, 98052, USA. Where he is the end-to-end architect for Azure Compute. Contact him at marcusfo@microsoft.com.

RICARDO BIANCHINI is a distinguished engineer with Microsoft Research, Redmond, WA, 98052, USA, where he leads the Systems Research group. He is a Fellow of ACM and IEEE. Contact him at ricardob@microsoft.com.