To appear in *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA '25)*

Chip Architectures Under Advanced Computing Sanctions*

August Ning
Princeton University
Electrical and Computer Engineering
Princeton, NJ, USA
aning@princeton.edu

David Wentzlaff
Princeton University
Electrical and Computer Engineering
Princeton, NJ, USA
wentzlaf@princeton.edu

Abstract

The rise of large scale machine learning models has generated unprecedented requirements and demand on computing hardware to enable these trillion parameter models. However, the importance of these bleeding-edge chips to the global economy, technological advancement, and strategic national interests have made them targets of sanctions. Recent advanced computing sanctions set limits on a device's Total Processing Performance, device bandwidth, and Performance Density and placed export controls on flagship data center and consumer products.

In this work, we present the first study on the architectural and *economic externality* implications of these advanced computing sanctions and their effects on large language model (LLM) inference. We identify which architectural parameters are limited under existing regulations, and perform thorough design space exploration of compliant designs. Optimized designs are able to improve LLM inference prefill performance by 4% and decoding performance by 27% compared to a restricted device baseline.

We then demonstrate how an architecture-first approach for computing policies allows chip designers and policymakers to craft efficient guidelines that achieve desired goals while minimizing negative externalities. We show how architectural features can unify marketing-based data center vs. non-data center regulations and how policies can be specified to create gaming-focused device architectures which are inherently limited in AI performance. Augmenting existing performance metrics with insightful architectural constraints better predict workload performance. Combined metrics achieved up to 42.4x narrower distributions compared to using theoretical compute performance alone, enable targeted and efficient policies.

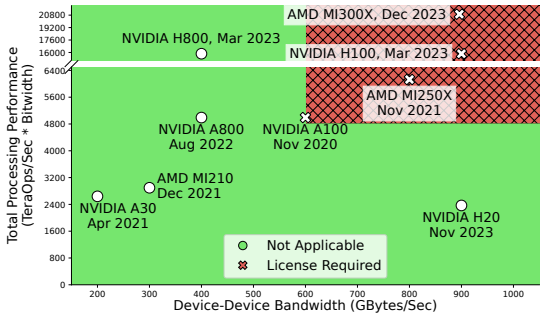
CCS Concepts

• **Computer systems organization** → **Architectures**; • **Social and professional topics** → **Import / export controls**; • **Computing methodologies** → *Artificial intelligence*.

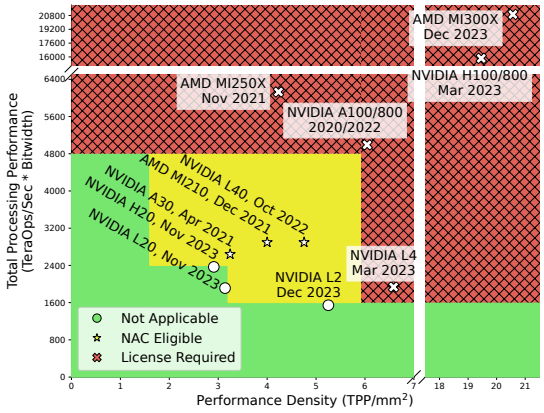
Keywords

Advanced Computing Sanctions, Artificial Intelligence

*The work in this paper has been included in August Ning's PhD thesis [48]. Text and figures are reproduced verbatim in both works. The authors of this paper are non-lawyer academic researchers. Nothing contained herein should be construed as legal advice or as a tool to ensure legal compliance.



(a) Device Classification Under October 2022 Specifications [14]. Inspired by [29].



(b) Device Classification Under October 2023 Specifications [16]. Inspired by [11].

Figure 1: Device Classification Under October 2022 and October 2023 Advance Computing Rule Specifications. Data from [1, 2, 24, 49, 52–54, 56, 67, 75]

ACM Reference Format:
August Ning and David Wentzlaff. 2025. Chip Architectures Under Advanced Computing Sanctions. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA '25), June 21–25, 2025, Tokyo, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3695053.3731012

ISCA '25 (to appear), Tokyo, Japan

ACM ISBN 979-8-4007-1261-6/2025/06
https://doi.org/10.1145/3695053.3731012

1 Introduction

Semiconductors and computer chips have become indispensable foundational technologies and securing performant computing has become key objectives for companies and governments alike. The prevalence of machine learning has fueled demand for graphics processing units, data center accelerators, and novel hardware developments to enable larger, more accurate, and more capable models.

The recent meteoric rise of large language models (LLMs) has again emphasized the importance of hardware. As LLM capabilities scale by increasing model sizes, state-of-the-art models are reaching billions and trillions of parameters [10, 25]. To train and provide ML and LLMs services, companies must use bleeding-edge chips in order to efficiently develop these massive models [2, 54].

Machine learning, LLMs, and computing hardware have become crucial tenets of technology companies as well as strategic government priorities worldwide. As firms aim to maintain competitive advantages, acquiring state of the art hardware has become a major bottleneck due to their complex manufacturing, vulnerable supply chains, and in turn high costs [32, 34, 35, 47, 79]. To protect strategic interests, governments have also imposed high profile sanctions on the devices used to train and run AI applications, making it difficult for some firms to develop their desired models.

Sanctions targeting advanced computing introduced in October 2022 placed export restrictions on chips that exceed performance based thresholds [14, 16]. Figure 1 shows how some devices have been classified under these regulations. These sanctions introduce performance limits which are disparate from how computer architects usually approach chip design. Traditionally, architects focus on optimizing performance within physical and cost constraints. However to comply with these export restrictions, performance and die area are now primary constraints and designers have modified designs to meet these new criteria in order to keep selling chips to fuel worldwide computing demand [37, 45]. Furthermore, although regulations focus on restricting flagship data center designs used for AI applications, updated regulations may require powerful "non-data center marketed" devices to also obtain export licenses. The licensing requirements are ultimately decided on a case-by-case basis which further complicates compliance [16].

Due to the recency of these regulations, manufacturers have focused on modifying existing product lines to create compliant designs [37, 45]. However, this often involves taking more powerful dies and disabling functionality until the device is under the export controls' thresholds. This is inefficient as it reduces the proportion of dies which are used for more profitable flagship devices and these performance-capped dies are larger compared to creating custom regulation-specific dies.

In this work, we present the first architectural exploration of LLM inference performance under these new sanctions. We demonstrate how chip architecture can be optimized within existing sanction definitions and improve on LLM prefill and decoding compared to a sanctioned device baseline. Additionally, we use our insights to propose architectural indicators which better predict performance for modern workloads compared to current metrics used for device classification.

Furthermore, we show how an architecture-first approach to hardware policy can reduce the economic *negative externalities* introduced by current regulations. By understanding the unique architectural feature and bottlenecks of different product segments and workloads respectively, policy specifications and hardware architectures can be domain-tailored to be inherently performance limited for workloads-of-interest. We show how policy can be scoped to create gaming-focused hardware which are architecturally limited in AI performance. This enables better targeted policies which reduce market distortions introduced by regulations.

Table 1: Advanced Computing Rule Definitions

(a) October 2022 Definitions [14]

Class.	All Devices
Regular License	TPP ≥ 4800 AND Bidirectional Device BW ≥ 600 GB/s

(b) October 2023 Definitions [16]

Class.	Data center	Non-data center
Regular License	TPP ≥ 4800 OR TPP ≥ 1600 AND PD ≥ 5.92	-
Notified Advanced Computing	4800 > TPP ≥ 2400 AND 5.92 > PD ≥ 1.6 OR TPP ≥ 1600 AND 5.92 > PD ≥ 3.2	TPP ≥ 4800

This work makes the following contributions:

- Detailed overview on which architectural components are limited by advanced computing rules.
- Thorough design space exploration of LLM inference chip architectures under these constraints including quantitative performance, die area, and cost analysis.
- Regulation compliant hardware optimizations that improve LLM inference prefill and decoding performance by up to 4% and 27% respectively compared to a modeled NVIDIA A100.
- Architectural performance indicators that better correlate with modern workload performance compared to existing metrics and reduce performance variation by up to 42.4x.
- An architecture-first approach which creates fine-grained and efficient policies that reduce negative externalities.

2 Background and Motivation

2.1 Advanced Computing Rule Sanctions

Chips and computing are often targeted by sanctions and export controls due to the importance of semiconductors within the global economy, strategic national interests, and concentrated supply chains. These sanctions often target specific computing applications and/or semiconductor manufacturing. For example, the United States has export controls regarding cryptography, including integrated circuits used for encryption [13]. The multilateral Wassenaar Arrangement has specific restrictions regarding electronics and computers that have dual-use applications [9]. The 2019-2023 Japan-South Korea trade dispute focused on restricting Japanese exports of semiconductor manufacturing chemicals to South Korea [39]. Japanese companies produce over 90% of the world's fluorinated polyimides and photoresists and account for over 92% of South Korean's supply of these chemicals, while South Korea manufactures 73% of the global DRAM and 51% of global NAND flash markets [30, 65].

In October 2022, the United States' Department of Commerce's Bureau of Industry and Security (BIS) introduced new export controls on advanced computing chips dubbed the Advanced Computing Rule (ACR), which applies to the following devices (referred to as "integrated circuits" in the regulations) **that can achieve an aggregate bidirectional I/O transfer rate over 600 Gbyte/s AND achieves aggregate *Total Processing Performance (TPP)* over 4800** [14]. The October 2022 ACR is also summarized in Table 1a.

Total Processing Performance (TPP) is defined as the theoretical maximum tera (10^{12}) operations per second (TOPS) multiplied

by the bitwidth of the operation. For devices that can operate on multiple bitwidths, TPP is determined by the max TOPS × bitwidth product. TPP is aggregated over all the dies within a package, such as chiplet devices. TPP is calculated based on non sparse operation performance. The guidelines also consider "tensor operations", which may combine a floating point multiply and accumulate as a single operations, as two operations when calculating TPP.

In October 2023, the advanced computing rule was modified to remove the I/O transfer rate restriction and introduced *Performance Density* (PD) thresholds. Performance density is defined as a device's TPP divided by *applicable die area* (measured in mm^2) [16]. Applicable die area only applies to dies in the device manufactured using a non-planar transistor architecture (*e.g.* sub 16nm FinFETs). The new rules differentiate between "data center" and "non-data center" designed/marketed devices - some data center and all non-data center devices can qualify for Notified Advanced Computing (NAC) license exceptions to allow exports with potentially fewer restrictions. The updated rules are described in Table 1b.

In December 2024, the BIS added new export controls on commodity high bandwidth memory (HBM) packages [17]. HBM packages with a "memory bandwidth density", which is defined as the memory bandwidth of the package divided by the package's area, greater than 2 GB/s/mm^2 are subject to this export control. Packages with a memory bandwidth density less than 3.3 GB/s/mm^2 may apply for license exception HBM which would allow exports to sanctioned countries if granted. This regulation **does not** apply to HBM which is installed inside computing devices before export. January 2025's proposed regulation added further licensing requirements for "front-end fabricators" and outsourced semiconductor assembly and test (OSAT) firms who manufacture export-controlled devices, as well as introduced new policies and licensing requirements that limit the quantity of AI-focused devices that can be exported to non-sanctioned countries [18, 19].

2.2 Sanctions' Effect On Chip Architecture

The October 2022 and October 2023 specifications target device performance and architectural features and chip designers have adapted their existing designs in order to keep selling their devices under the new regulations. Currently, the October 2023 specification are still in affect, as December 2024 and proposed January 2025 updates did not change device-level export controls. In October 2022, the flagship data center GPUs were the NVIDIA H100 (announced March 2022) and AMD MI250X (launched November 2021) [1, 72]. The NVIDIA H100 [54] has a TPP of 15824 and device bandwidth of 900 GB/s. The AMD MI250X [1] has a TPP of 6128 and device bandwidth of 800 GB/s. The NVIDIA A100 [53] was the flagship device available in October 2022, and has a TPP of 4992 and device bandwidth of 600 GB/s. The October 2022 specifications generally only applies to powerful flagship devices.

Firms have modified their flagship products to comply with restrictions. The October 2022 definitions do not apply to devices with either TPP < 4800 **or** device bandwidth < 600 GB/s, so manufacturers only had to reduce one parameter to comply with the regulations. The NVIDIA A800 [52] (released in August 2022) uses the same GA100 die as the sanctioned A100, but reduces the device bandwidth to 400 GB/s while maintaining the same 4992 TPP. The NVIDIA H800 [56] similarly has 15824 TPP and 400 GB/s which is based on the sanctioned H100.

The October 2023 updates and new PD requirements now sanction the previously regulation-specific devices such as the NVIDIA A800 (PD 6.04) and H800 (PD 19.45). The AMD MI210 data center GPU [1] (2896 TPP, 300 GB/s, PD 3.76) was previously unregulated, but now requires NAC exception for export. Similarly, with the new distinction between data center and non data center devices, NVIDIA's RTX 4090 gaming GPU [55] (5285 TPP, 32 GB/s, 8.68 PD) also now requires NAC exceptions.

Firms have again adapted to new regulations. NVIDIA in November 2023 announced new devices: H20, L20, and L2 which will comply with the updated October 2023 data center devices regulations [45]. NVIDIA also launched the RTX 4090D [50] (4708 TPP) which is based on the same AD102 die as the RTX 4090 but disables more compute cores (114 vs 128) to avoid the 4800 TPP threshold for non data center chips. AMD has also attempted to create sanction specific devices but have faced regulatory approval issues. The AMD's regulation-specific MI309 was denied export approval under the October 2023 specifications and it is unknown why the device was not approved [38].

2.3 Large Die Area GPU Designs

Recent flagship data center device dies have reach the near physical limit of die sizes or have used chiplets to achieve 1000+ mm^2 devices. As chip designers strive to improve single device performance and transistor scaling plateaus, total die area has increased. Larger dies are prone to manufacturing defects and reduced die yields; flagship devices are often fabricated on bleeding edge process nodes, which are less mature and more prone to defects. There are also physical limits of manufacturing, and current EUV technology limits single dies to around 860mm^2 [63]. Furthermore, fewer larger dies fit onto a single wafer and firms will need to order more wafers, increasing costs and manufacturing times compared to smaller dies [47]. These two effects compound and make manufacturing large chips expensive.

To ameliorate these design challenges, firms have turned to binning and chiplet designs to create multiple product lines from the same dies - both these tactics are motivated by cost. Binning allows partially defective chips to be salvaged to be reused in less powerful products, and chiplet devices split designs across multiple smaller dies which improves overall product yield compared to monolithic designs.

Both versions of the advanced computing rules have influenced how firms approach large die area designs. NVIDIA A800 and H800 devices uses the same dies as the A100 and H100 respectively and could be made from partially defective dies where the device bandwidth performance did not meet the 100 series' specifications or intentionally disabled to comply with regulations [37, 46]. TPP is calculated based on all the dies on the device, but chiplet based devices can reduce TPP or device bandwidth by reducing compute chiplet or IO chiplet counts respectively.

For the October 2023 definitions, large monolithic dies can decrease TPP and performance density by disabling compute cores; this was done for the NVIDIA H20 [45]. For chiplet based devices,

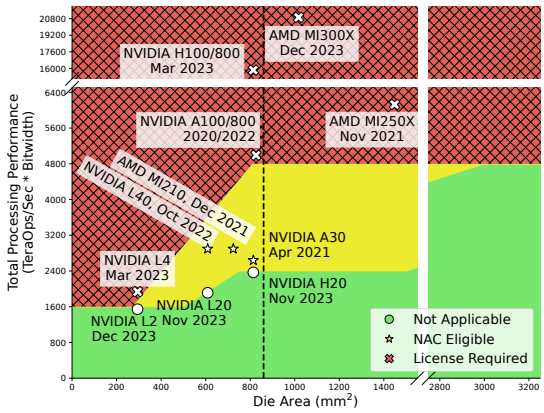


Figure 2: Device Die Area and Total Processing Performance Under October 2023 Specifications [16] - Devices can avoid ACR by increasing die area. Data from [24, 67, 75]

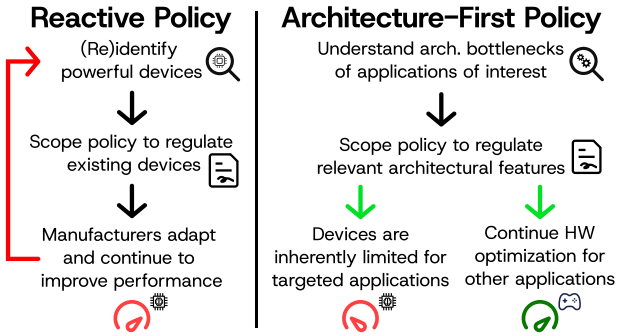


Figure 3: Architecture-First Policy Overview. Only regulating relevant architectural features creates fine-grained, efficient policies.

removing chiplets may reduce TPP, but may not reduce performance density since die area decreases as well. For chiplet designs to comply with existing PD restrictions, devices may need to disable computing cores within chiplets, which opposes the original motivation of using smaller chiplet designs to reduce die defects.

2.4 Market Distortions and Negative Externalities

In terms of economics, sanctions introduce *market distortion*, which is any interference between buyers and sellers where prices no longer reflect free-market conditions [26]. Current sanctions essentially reduce the supply of computing devices available, which in turn increases prices for companies who want to buy these chips. This market condition where supply and demand are artificially imbalanced is known as *deadweight loss* [44].

The October 2023 ACRs created separate classification guidelines between data center and non-data center devices. Although the updates ultimately required more devices to acquire licensing, the data center vs non-data center distinction recognizes that not all high performance devices are designed for the AI applications the sanctions are targeting [15]. These updated definitions introduced *negative externalities*: actions taken by one party causing indirect adverse effects on uninvolved third parties [44]. Regulations targeting powerful AI devices also required powerful gaming focused devices to acquire export licenses. The updated regulations reduced the global availability of data center and non-data center devices, increasing overall deadweight loss.

2.5 Motivation

The ACRs' restrictions are seemingly counter-intuitive to how computer architects may approach chip design. Traditionally, architects optimize their designs for power, performance, and area - either targeting/maximizing performance while minimizing device power and area. Moreover, the regulations were changed only a year after the initial specifications were announced. Chip design cycles can span multiple years and sudden regulation changes make it difficult and expensive for device manufacturers to adapt [3].

ACRs essentially add ceilings on theoretical device performance, as TPP sets limits on maximum operations per second, and practically only regulates "AI compute performance" such as NVIDIA Tensor Cores and AMD Matrix Cores. These cores are generally not crucial for gaming performance. Furthermore, the PD metric acts as a floor on die area. Figure 2 shows the October 2023 specifications based on die area rather than performance density. By keeping TPP constant and increasing die area, devices can avoid ACR export restrictions. For the October 2023 definitions, in order for a device with 2399 TPP to avoid the restrictions, the device needs to have a die area greater than 750 mm^2. For a 1600 TPP device to be NAC eligible, it needs to have a die area greater than 270 mm^2.

For a 4799 TPP design to avoid export restrictions, the device must have total die area greater than 3000 mm^2, which is more than three times greater than the current reticle limit. Compliant designs must be multi-chip module designs, which add an additional design space for die sizing to optimize performance as well as cost.

These theoretical performance limits may not reflect actual workload performance, as they focus on the compute component of a system and not the rest of the architecture. Furthermore, these metrics stem from compute regulations from the 1990s, and modern computers, workloads, and export control motivations have changed drastically in the past 30 years [8]. By only specifying TPP, device bandwidth, and performance density limits, there remains a large design space where designers can improve performance. In this work, we conduct a thorough architectural design space study on how chips can be optimized under the existing specifications. By providing detailed breakdowns on how key workloads are affected by current sanctions, computer architects can reason about which architectural improvements to focus on while also following current and future export control regulations.

Additionally, although sanctions target computing hardware, regulations are scoped to curtail specific workloads. However, the devices covered by the sanctions are used for multiple workloads with different architectural bottlenecks. For example, GPUs are often used in AI and stockpile stewardship applications, but are also used for gaming and weather forecasting [36, 57]. We use chip architecture insights to develop modern architecture-first performance predictors and propose efficient policy which only affects the product segment or workload-of-interest, as shown in Figure 3.

Governments and researchers globally have proposed new policies which may regulate computing hardware with respect to sustainability [27, 66], AI safety [60], cryptocurrency mining [31], *etc.* Moreover, the mercurial nature of politics contrasts multi-year long semiconductor design cycles, and suddenly regulation changes may further disrupt product roadmaps and profits for hardware manufacturers. Our proposed architecture-first approach for hardware policy sets a framework where computer architects work with policy makers to scope regulations to affect only the most relevant architectural features and manufacturers can continue to sell optimized devices for non-target workloads, which increases revenue and reduces negative externalities.

3 Methodology

In this section, we provide background on large language models and LLM inference performance metrics, the LLMCompass [82] evaluation framework we use for architectural design space exploration, and overview how we interpret the advanced computing rules and their effects on chip architecture.

3.1 LLMs and Performance Metrics

Although artificial intelligence applies to a wide range of workloads, LLMs have become the leading model architecture. LLMs are models with a large amount of parameters that have been pre-trained on large corpora of data. In this work, we focus on Decoder-only Transformer models, which are the most popular variant and are adopted by LLaMA [28, 71], GPT-3 [10], PaLM [22], *etc.* LLMs are comprised of stacks of identical Transformer layers.

LLM inference can be divided into two separate phases: (1) Prefill - after receiving the input prompt, all the input tokens are processed in parallel to generate the first token and the KV cache. (2) Decoding - afterwards, the output tokens are generated one by one in an auto-regressive manner. Two key performance metrics for LLM inference are (1) time to first token (TTFT) - the latency of the prefill stage and (2) time between tokens (TBT) - the per-token latency of the decoding stage. TTFT and TBT can be used to derive performance metrics such as end-to-end latency and throughput.

A common metric used for evaluating LLM hardware performance is model FLOPs utilization (MFU), defined as the ratio of the observed throughput relative to the theoretical maximum throughput of a system operating at peak FLOPs [22]. Related work [82] shows that LLM inference can achieve near peak theoretical FLOPs during the compute-intensive prefill stage but suffer from low utilization during the memory-intensive decoding stage. LLM operations such as Softmax, LayerNorm, and GeLU have low arithmetic intensities and cannot achieve high throughput during inference [81].

3.2 LLMCompass Framework

In this work, we use LLMCompass to explore how different hardware designs affect LLM inference [82]. LLMCompass is a high-level hardware simulation framework tailored for LLM workloads. Figure 4 shows the hardware template used by LLMCompass. Each device has multiple cores, a shared global buffer between cores, which is connected to the off-chip memory and device-device interconnect. Each core can have multiple lanes sharing a local buffer. Each lane is composed of a vector unit and a systolic array.

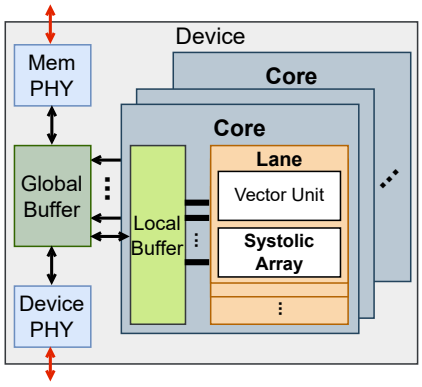


Figure 4: LLMCompass' Hardware Template. Systolic array size, lane count, and cores per device configurations determine TPP.

Table 2: Model Architectures

Parameter	GPT-3 175B [10]	LLaMA 3 8B [28]
Number of Layers	96	32
Model Dimension	12288	4096
FFN Dimension	49152	14336
Attention Heads	96	32
K/V Heads	96	8
Activation Function	GELU	SwiGLU

We configure LLMCompass to evaluate the computational patterns of two models: GPT-3 175B [10] and Llama 3 8B [28]. The model architectures are summarized in Table 2. To model Llama 3 8B, we used a modified version of LLMCompass which supports grouped-query attention [5] and SwiGLU activation [61]. As LLMs are composed of stacks of repeated transformer layers, we only need to simulate and report results for one layer. For each model, we simulate a standard layer with batch size 32, input sequence length 2048, and output sequence 1024, which is a typical setting for LLM inference workloads ran on flagship data center GPUs.

3.3 Interpreting The Specifications

The ACRs specify limits on TPP and device bandwidth/performance density. While device bandwidth and PD are straightforward to calculate, TPP requires more nuance to translate to chip architecture. TPP is calculated using peak theoretical performance reported by device manufacturers. Although the exact way these theoretical peaks are derived is unknown, they can generally be calculated by multiplying the total number of operations a device can compute in a single cycle by the device clock frequency. GPU manufacturers report tensor/matrix compute theoretical peaks separately from their vector compute peaks, as they achieve higher performance which is used when calculating TPP. With LLMCompass, we calculate TPP based on the systolic array's configuration.

Each systolic array can compute $DIM_X * DIM_Y$ MAC-OPs/cycle and each multiply-and-accumulate is counted as two FLOPs. We configure LLMCompass such that each lane has one systolic array, and each core can have multiple lanes. We use the NVIDIA A100's

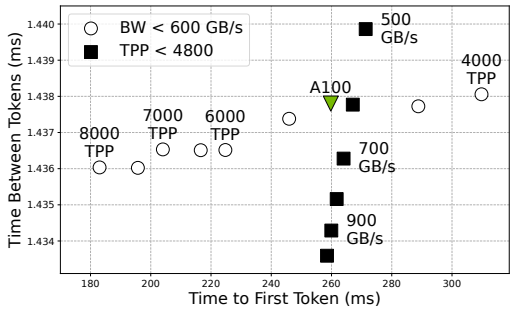


Figure 5: Prefill and Decoding Latency Modeling GPT-3 175B, Sweeping TPP or Device Bandwidth for October 2022 Specifications. White circle markers have device bandwidth < 600 GB/s and black square markers have TPP < 4800. Green triangle denotes modeled A100.

1410 MHz clock frequency and use FP16 systolic arrays to calculate the maximum FP16 units that can be used for systolic arrays in single device.

For a GPU-like device with systolic arrays to be under a given TPP, the configuration must meet the following:

$$FP_{max}(TPP) \geq DIM_X * DIM_Y * LC * CD \quad (1)$$

where $FP_{max}(TPP)$ is the maximum number of systolic array FPUs for a given TPP and device clock frequency, DIM_X * DIM_Y are the dimensions of the systolic arrays, LC is the lanes per core, and CD is cores per device. In our experiments, we sweep systolic array dimensions and lanes per core count and change cores per device accordingly to keep design points within TPP targets. To limit device bandwidth, we change the device PHY counts and per PHY bandwidth in LLMCompass. To calculate performance density, we use LLMCompass' area and cost model to find the design's die area and silicon costs. These estimates are based on the 7nm process which is the same process used by the NVIDIA A100's dies.

4 Chip Architecture Optimization Under Advanced Computing Rules

In this section, we perform design space explorations to evaluate the architectural implications of current regulations and demonstrate how chips can be optimized for LLM inference under advanced computing rules. Reported latency results in this section and Sec. 5.3 are simulated using LLMCompass, including results for the modeled NVIDIA A100. Die area results in this section and Sec. 5.3 come from LLMCompass but we use the GA100 [53] die area for the modeled A100.

4.1 Oct. 2022 - TPP vs Bandwidth Scaling

Under the October 2022 specifications, devices can avoid ACR restrictions if they have either TPP < 4800 **or** device bandwidth < 600 GB/s. These definitions allow device manufacturers to continue to scale one of these "knobs". Between October 2022 to October 2023, regulation specific devices have capped device bandwidth while increasing TPP (NVIDIA A800, H800), while recent devices have capped TPP with increased device bandwidth (NVIDIA H20).

To explore the trade offs of TPP vs device bandwidth scaling, we configure LLMCompass based on the NVIDIA A100's architecture.

Table 3: Design Space Exploration Parameters Compared to A100 (GA100)

Parameter	A100 (GA100) [53]	DSE
Core Count	108 (128)	-
Systolic Array Dim.	16x16	16x16, 32x32
Lanes per Core	4	1, 2, 4, 8
Private L1 Cache (KB)	192	192, 256, 512, 1024
Shared L2 Cache (MB)	40 (48)	32, 48, 64, 80
HBM Mem. Capacity (GB)	40/80	80
HBM Bandwidth (TB/s)	2 (2.4)	2, 2.4, 2.8, 3.2
Device Bandwidth (GB/s)	600	Fig. 6: 600 Fig. 7: 500, 700, 900

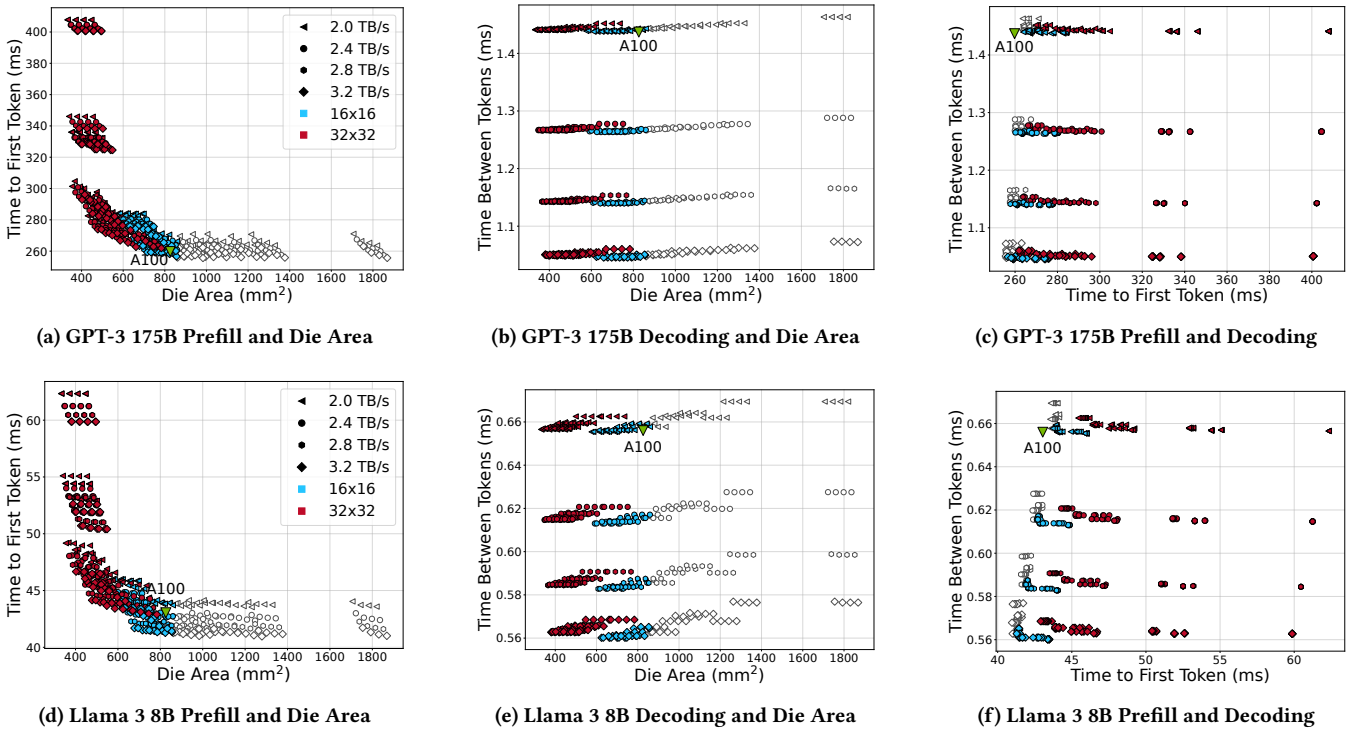
For the devices with capped TPP < 4800, we set device core count to 103 (TPP 4759) and sweep device-to-device PHY count. Similarly, we set devices with capped device bandwidth < 600 GB/s by reducing per device-to-device PHY bandwidth and sweep core count. Figure 5 shows the TTFT and TBT results of these two parameter sweeps simulating GPT-3 175B. White circle markers are configurations with fixed device bandwidth and labeled TPP and black square markers have fixed TPP and labeled device bandwidth. All configurations are not regulated by the October 2022 ACR except the NVIDIA A100.

The results show that increasing TPP/core count is better for reducing TTFT latency and increasing device bandwidth is better for reducing TBT latency. This is in line with expectations, as the prefill stage is generally compute bound while decoding stage is bandwidth bound. Although decoding latency is more sensitive to device bandwidth, the effect is minimal - increasing device bandwidth from 600 GB/s to 1000 GB/s only decreases TBT by 0.27%. Increasing TPP to decrease prefill latency is much more rewarding. Increasing TPP from 4000 to 5000 decreases TTFT by 16.2%.

Increasing TPP or device bandwidth also increases die area. The 7000 TPP configuration decreases TTFT by 34.1% compared to 4000 TPP, but the die area also increases by 48.3%. With a die area of 854 mm^2, the 7000 TPP design is at the reticle limit and is unlikely for all cores to be fully functional. In summary, the October 2022 specifications allow prefill latency improvements by scaling TPP/core count but may be capped by die area constraints. The specifications also prevent significant decoding latency improvements when only scaling TPP or device bandwidth.

4.2 Oct. 2022 - Design Space Exploration

Although the October 2022 specifications only set limits on TPP and device bandwidth, there are many architectural parameters that can still be changed which affect workload performance while complying with ACRs. As previously discussed in Section 3.3, TPP limits total systolic array FPU count, but does not further regulate how these systolic arrays are configured. Moreover, there are no limits on how the memory system is configured. We study this design space by configuring design point s to have TPP near 4800 and device bandwidth = 600 GB/s and sweep architectural parameters as show in Table 3 (512 total designs). LLM inference performance results are shown in Figure 6. The NVIDIA A100 is again shown for reference but does not comply with ACRs. Many design points have die areas larger than the reticle limit - designs with die areas larger than 860mm^2 have white markers.



(a) GPT-3 175B Prefill and Die Area

(b) GPT-3 175B Decoding and Die Area

(c) GPT-3 175B Prefill and Decoding

(d) Llama 3 8B Prefill and Die Area

(e) Llama 3 8B Decoding and Die Area

(f) Llama 3 8B Prefill and Decoding

Figure 6: Prefill, Decoding, and Die Area for 4800 TPP, 600 Device BW Design Space Exploration Modeling GPT-3 175B and Llama 3 8B. Marker shape indicates the design point's memory bandwidth. Marker color indicates the design point's systolic array configuration. White markers violate the 860mm^2 reticle limit. Green triangle denotes modeled A100.

Prefill is compute bound which makes it difficult to decrease TTFT latency while TPP is limited. However, design points exist where October 2022 ACR compliant designs have lower TTFT compared to the A100. Numerous design points are able to achieve lower TBT latency, as decoding is memory bandwidth bound. Figures 6b and 6e show clear levels of decoding performance grouped by memory bandwidth (shown by marker shape).

Figures 6c and 6f show that there are ACR compliant designs which improve on prefill and decoding latency compared to the A100, but many of these designs have die areas larger than the reticle limit. When only considering manufacturable single die designs, GPT-3's optimized design decreases TTFT by 1.2% and TBT by 27% compared to an A100 baseline. This configuration is similar to the A100, but decreases lanes per core to 2, increases L2 cache to 64 MB, and maximizes memory bandwidth to 3.2 TB/s. The Llama 3 optimized design decreases TTFT by 4% and TBT by 14.2%. This configuration differs from the A100 with 512 KB L1 caches, 64 MB L2 cache, and 3.2 TB/s memory bandwidth.

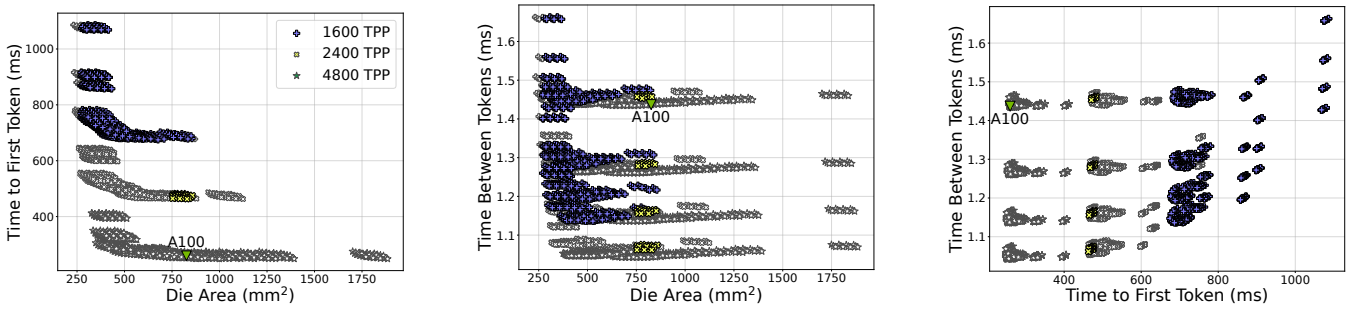
Decreasing lane count or increase L1 cache size increases the effective private cache size per systolic array, increasing L2 cache size helps with the compute bound prefill stage, and increasing memory bandwidth significantly improves decoding performance. GPT-3's and Llama 3's designs' die areas are 856 mm^2 and 823 mm^2 respectively. GPT-3's configuration is on the edge of manufacturability, but shows how chip architectures can continue to improve LLM inference performance under October 2022 ACRs.

4.3 Oct. 2023 - Design Space Exploration

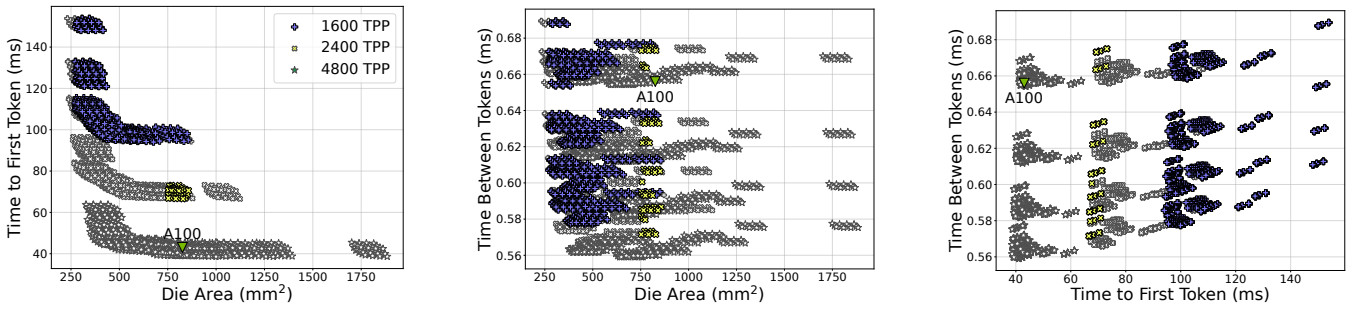
The October 2023 ACR updates removed the device bandwidth restrictions, introduced performance density, and created new tiers of restrictions depending on TPP and performance density. As previously discussed, PD adds a minimum die area requirement for ACR compliant designs. Increasing die area may improve workload performance but may lead to diminishing returns, especially when TPP is also limited. moreover, the decreased die yield and increased silicon cost may eventually outweigh the performance improvements.

We perform a large design space exploration for 1600, 2400, and 4800 TPP designs under the October 2023 specifications. The sweeped architectural parameters are shown in Table 3 (1536 designs per TPP). These selected TPPs are used in the October 2023 ACR definitions, each with different performance density cut offs. We classify designs based on the performance density to not be regulated by the ACRs, as NAC eligible devices may not always be granted export licenses [38]. LLM inference performance results are shown in Figure 7. Marker shape and color indicate TPP, and white markers indicate designs that either violate performance density or reticle limits.

The low performance density requirement make all 4800 TPP designs invalid and reducing TTFT latency becomes impractical. Even for compliant 2400 TPP designs, the fastest TTFT is still 78.8% and 54.6% slower compared to the A100's for GPT-3 and Llama 3 respectively. However, decoding performance can still be improved



(a) GPT-3 175B Prefill and Die Area (b) GPT-3 175B Decoding and Die Area (c) GPT-3 175B Prefill and Decoding



(d) Llama 3 8B Prefill and Die Area (e) Llama 3 8B Decoding and Die Area (f) Llama 3 8B Prefill and Die Area

Figure 7: Prefill, Decoding, and Die Area for 1600, 2400, and 4800 TPP Devices Design Space Exploration Modeling GPT-3 175B and Llama 3 8B. White markers indicate devices that violate PD limits or the 860mm^2 reticle limit. Green triangle denotes modeled A100.

Table 4: GPT-3 175B Performance Density Compliant Designs Optimal Design Comparison

Parameter	PD Compliant	Non-Compliant
Die Area	753 mm^2	523 mm^2
PD	3.18	4.59
TTFT	465 ms	470 ms
TBT	1.062 ms	1.053 ms
Silicon Die Cost (7nm)	$134	$88
1M Good Dies Cost (7nm)	$350M	$177M

because memory bandwidth is not regulated. The fastest TBT for 1600 and 2400 TPP designs running GPT-3 are 20.9% and 26.1% faster respectively compared to the A100's; the fastest TBT Llama 3 designs are 12.0% and 12.8% faster. Therefore, the October 2023 specifications are more effective at preventing prefill performance improvements, but still allow decoding improvements.

4.4 Oct. 2023 - Performance Density and Cost

Performance density restrictions for 2400 TPP designs and the reticle limit only allows a 110 mm^2 area budget range for single die designs, even though it is possible to achieve similar performance with less die area. From the 1536 design points, there are only 56 valid 2400 TPP designs - 1429 designs violate performance density and 51 violate the reticle limit.

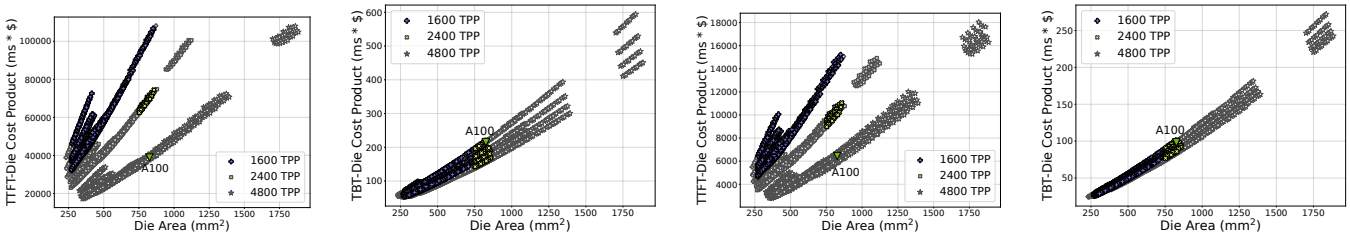
Table 4 shows the fastest TTFT design for GPT-3 for PD compliant and non-compliant 2400 TPP designs. Despite both designs having similar performance, the PD compliant design is 44% larger and silicon costs are 52.3% higher compared to the non-compliant design. Factoring in die yield, the cost for 1 million good dies is almost double the cost for the PD compliant design.

Minimum die area requirements also increase device power. Comparing the two devices, they have identical architectures except for cache configuration, where the PD compliant device has 1 MB L1 cache and 48 MB L2 cache, while the non-compliant device has 192 KB L1 and 32 MB L2. The PD compliant device has almost triple the floor planned SRAM area (151 MB vs 52 MB of on chip SRAM). If all are turned on, these caches increases static and dynamic power which increase operating costs.

Figure 8 shows the latency-die cost product (lower is better for both parameters) for the design space exploration from Figure 7. The October 2023 specifications prevent the most prefill latency-cost efficient 4800 TPP designs from being exported. Furthermore, the 2400 TPP designs are again significantly impacted by performance density requirements. GPT-3's PD-compliant, minimum latency-cost designs have 2.72x and 2.64x higher prefill and decoding latency-cost product respectively compared non-compliant designs. For Llama 3, compliant designs have 2.58x and 2.91x higher prefill and decoding latency-cost products respectively.

In summary, this section has the following takeaways:

- ACRs are effective at preventing prefill latency improvements but still allow decoding improvements.
- The reticle limit prevents many single die compliant designs from begin manufactured.



(a) GPT-3 175B TTFT×Cost (b) GPT-3 175B TBT×Cost (c) Llama 3 8B TTFT×Cost (d) Llama 3 8B TBT×Cost

Figure 8: TTFT and TBT Latency-Cost Product For 1600, 2400, and 4800 TPP Devices Design Space Exploration. Lower is better. White markers indicate devices that violate PD limits or the 860mm^2 reticle limit. Green triangle denotes modeled A100.

- The performance density requirement does not correlate well with performance and increases costs.

5 Efficient Architecture-First Policy

This section overviews how current advanced computing rule specifications increased negative economic externalities. Currently, advanced computing sanctions may be reactively designed, where policy makers and manufacturers are caught in a "cat-and-mouse" game between updating regulations and improving performance. We show how an architecture-first approach guides designers towards domain-tailored hardware which promotes efficient policy and reduces deadweight loss.

5.1 Existing Policy's Negative Externality

The October 2023 updates introduced separate classifications for data center and non-data center devices and NAC license exceptions for qualifying devices. One of the stated goals of the advanced computing rules is to prevent sanctioned entities from receiving AI focused hardware [62]. Although broadly defined, existing ACRs are *effective*, since they can identify flagship GPU devices used for bleeding-edge AI applications such as LLMs. However, these specifications also restrict devices **not** designed for AI such as top-of-the-line gaming GPU devices. This is a negative externality because this policy scoped for preventing AI applications has (potentially) inadvertently also prevented gaming applications. Furthermore, currently ACRs are not *economical*, as creating compliant devices requires additional costs compared to some sanctioned devices as discussed in Section 4.4. By understanding these externalities, policies can be scoped to only apply to relevant devices while being intuitive and economical for device manufacturers to follow.

5.2 Marketing-Based Classification

The October 2023 specifications apply different regulations for data center devices. A downside of this classification is that the difference between a data center and non-data center device is based on marketing, and a single die architecture may be used in both data center and non-data center marketed devices. Under October 2023 specifications, some regulated data center devices can avoid regulations if they were rebranded as consumer devices. Similarly, some non-data center devices would be restricted if they were evaluated under the stricter data center device guidelines.

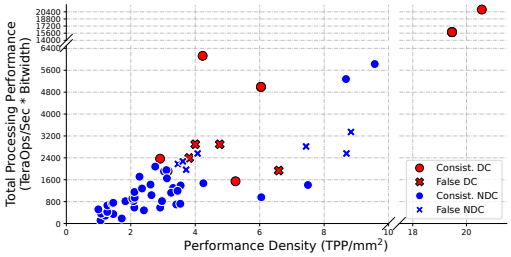


Figure 9: October 2023 Marketing-Based Device Scatter Plot. "False" devices have differing regulations if they were rebranded as the opposite market segment. Data from [40, 41, 67, 75].

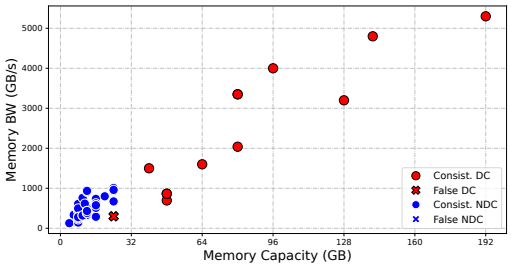


Figure 10: Device Scatter Plot Based on Memory Capacity and Memory Bandwidth. Architecture-based classification has no false non-data center and only two false data center devices. Data from [40, 41, 67, 75].

To investigate, we calculated TPP and PD for 65 GPUs released by AMD and NVIDIA between 2018 and 2024; 14 devices are marketed as data center devices, and 51 are marketed as consumer or workstation devices. We classified each device based on data center and non-data center specifications. A device is considered consistently classified if it is unregulated or regulated for both specifications. We classify a device as a "false data center" device if a data center marketed device is currently regulated, but would not be regulated if it was marketed as a consumer device. Similarly, a "false non-data center" device is a non-data center marketed device which is currently not regulated, but would be regulated if it was marketed as a data center device. These classifications are plotted in Figure 9. Existing specifications result in 4 false data center devices and 7 false non-data center devices.

Flagship gaming GPUs such as the NVIDIA RTX 4080 and AMD RX 7900 XTX would be regulated if they were marketed as data

center devices. Furthermore, Low TPP data center devices such as the NVIDIA L40 and A40 would not be restricted if they were instead marketed as workstation devices. These results demonstrate the flaws of using only marketing-based differentiation; existing policies incentivize manufacturers to market powerful consumer devices to face fewer restrictions. This also introduces inefficiencies for data center operators, as some manufacturers prevent deploying consumer devices within data centers [51].

We can address these discrepancies by applying the same regulations across all devices and leverage architectural metrics to create the desired classification. Data center GPUs have different target users, workloads, and operating environments compared to consumer devices, so their architectures generally have higher memory capacity and higher memory bandwidth. We can differentiate current data center devices by identifying devices with more than 32 GB memory **or** more than 1600 GB/s memory bandwidth, which is show in Figure 10. This classification results in no false non-data center and only two false data center devices. The two false data center devices, the NVIDIA L2 and L4 GPUs, are based on the AD104 die which is also used in high end gaming GPUs [70]. Using architectural metrics for hardware policy reduces inconsistencies from marketing-based differentiation, and provides clear guidelines for manufacturers on how to scope potential future architectures if they need to comply with said policies.

5.3 Architecture-First Performance Indicators

As shown in Section 4, current ACRs still permit a broad range of architectural configurations with varying performance. Figure 11 shows the TTFT and TBT distributions for all 4800 TPP configurations with die area's within the reticle limit from Figure 7. The first column shows all 4800 TPP configurations and following columns show distributions with a single fixed architectural parameter.

Even for the two stages of transformer inference, the fixed architecture distributions vary. Compute bound TTFT performance benefits from having more L1 cache per systolic array. Designs which only have 1 lane per core (Figures 11a, 11c column 2) have 5x and 3.3x narrower distributions respectively for GPT-3 and Llama 3 compared to TPP alone. As previously discussed, TBT performance is significantly affected by memory bandwidth - designs with fixed 2.8 TB/s memory bandwidth (Figures 11b, 11d column 5) have a 20.6x (GPT-3) and 10.7x (Llama 3) narrower distribution.

The October 2022 ACR had a minimum device bandwidth restriction, but device bandwidth does not correspond well with LLM inference performance - configurations with fixed 500 GB/s device bandwidth have only 5.7% (GPT-3), 15.2% (Llama 3) smaller TTFT distributions compared to limiting TPP alone, as well as negligible effects on TBT distributions. **Narrow distributions indicate strong performance correlation which can be used to efficiently target workloads-of-interest.**

We now demonstrate how architectural parameters can be used to improve computing policy. Regulators have been interested in limiting device performance, so we perform another design space exploration where architectural parameters are decreased compared to a modeled NVIDIA A100. We again set TPP to 4800 and filter out designs with die areas greater than the reticle limit. The parameters are shown in Table 5, resulting in 2304 configurations and performance distributions are shown in Figure 12.

Table 5: Design Space Exploration Parameters For Figure 12. Bold parameters are same as A100 [53].

Parameter	DSE
Systolic Array Dim.	4x4, 8x8, **16x16**
Lanes per Core	1, 2, **4**, 8
Private L1 Cache (KB)	32, 64, 128, **192**
Shared L2 Cache (MB)	8, 16, 32, **40**
HBM Bandwidth (TB/s)	0.8, 1.2, 1.6, **2**
Device Bandwidth (GB/s)	400, 500, **600**

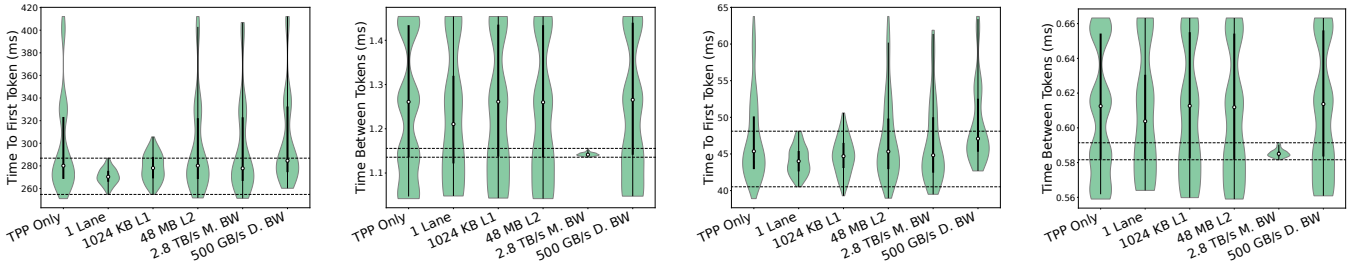
When targeting TTFT performance, restricting L1 cache size has the slowest median TTFT and narrowest performance distribution. Small L1 caches slow down data provisioning to the systolic arrays and become the major performance bottleneck. Devices with 32 KB L1 caches have a median TTFT 58.7% (GPT3), 52.6% (Llama 3) slower compared to a modeled A100. 32 KB L1 cache devices have 1.59x (GPT-3), 1.43x (Llama 3) narrower distribution respectively compared to restricting TPP alone. Limiting memory bandwidth significantly increases TBT latency. Devices with 800 GB/s memory bandwidth have median TBT 110% (GPT-3), 58.7% (Llama 3) slower compared to a modeled A100 and have 41.8x (GPT-3), 42.4x (Llama 3) narrower distributions.

Architecture-first constraints can directly target the workload's unique bottlenecks and serve as better performance indicators for modern workloads compared to using theoretical performance alone. The narrower performance distributions provides better control over expected performance for future chip architectures and policy. Further, they allow finer-grain disambiguation of a single workload: future policy can target TTFT and TBT performance separately using this architecture-first approach. TTFT or TBT performance can be limited by reducing L1 cache size or memory bandwidth respectively, and limiting these parameters independently will permit designs that do not affect the other stage's performance.

5.4 Externality-Aware Policy

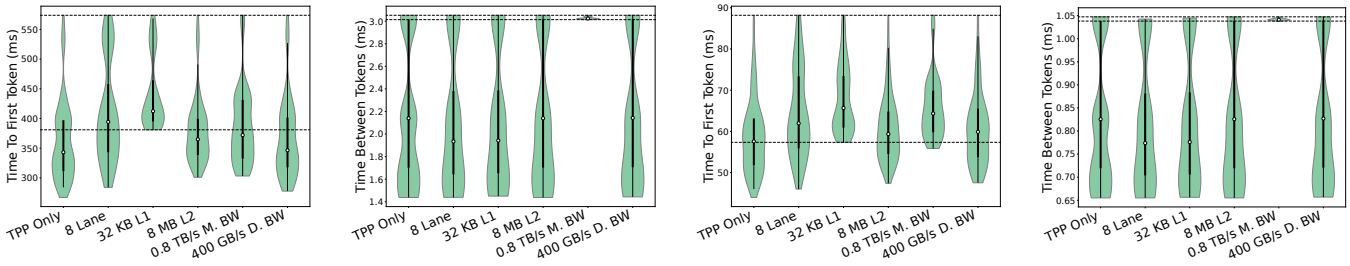
Using architectural performance indicators allows device designers and regulators to work together and craft efficient policies. From our results and existing literature, we show how architecture-first policies can achieve their goals while minimizing negative externalities as well as potential trade-offs. As a case study, we show how policy can be scoped to restrict device architectures which excel in AI applications but allows architectures which explicitly achieve high gaming performance. Following existing regulations, we suggest architectural parameters that are commonly disclosed on device datasheets and white papers.

Matrix Multiplication Performance: Matmul hardware such as systolic arrays are crucial for achieving high performance matrix multiply operations used in machine learning applications. However, gaming applications rely on the GPU's SIMT architecture, texture units, and ray tracing cores to speed up graphics rendering. Modern consumer GPUs (such as the NVIDIA RTX 4090) have systolic arrays to provide versatility, but this has caused these devices to be restricted under current guidelines. If policies were scoped to restrict device matmul performance (*e.g.* limit tensor TFLOPs for



(a) GPT-3 175B TTFT (b) GPT-3 175B TBT (c) Llama 3 8B TTFT (d) Llama 3 8B TBT

Figure 11: TTFT and TBT Latency Distributions for 4800 TPP From Fig. 7 DSE Grouped by Select Architectural Parameters. The first column includes all configurations and following columns show distributions with a single fixed architectural parameter.



(a) GPT-3 175B TTFT (b) GPT-3 175B TBT (c) Llama 3 8B TTFT (d) Llama 3 8B TBT

Figure 12: TTFT and TBT Latency Distributions for Table 5 DSE Grouped by Select Architectural Parameters. The first column includes all configurations and following columns show distributions with a single fixed architectural parameter.

NVIDIA devices or remove systolic arrays entirely), gaming-focused architectures would be encouraged to be redesigned without matmul accelerators which would hamper their AI performance but would likely maintain high gaming performance [55].

Systolic arrays have been used in graphics upscaling technologies such as NVIDIA's DLSS [20]. However, other compatible upscaling technologies are available which can run on GPUs which do not have systolic arrays [6]. If systolic arrays are necessary for gaming focused designs, policies may regulate the array's dimensions, as smaller systolic arrays perform worse in LLM inference applications [82].

On-Chip SRAM Memory Sizing: We previously showed how L1 cache sizing has the most direct impact on limiting LLM inference TTFT performance, but gaming applications are also sensitive to cache sizing to mitigate irregular memory accesses and latency [42]. Additionally, there needs to be nuance on how SRAM-related specifications are crafted as manufacturers vary on how they classify on-chip SRAM. Google TPU's common memory acts as a global L2 between the systolic arrays [33] and NVIDIA's Shared Memory combines private, per-SM scratchpads with L1 cache [53].

Nonetheless, cache configurations are a promising architectural performance indicator, and cache sizing based policy can still be implemented to differentiate between data center and non-data center devices as they already have different cache hierarchies. NVIDIA's Hopper data center architecture has 256 KB of L1 cache/shared memory per SM while their Ada Lovelace consumer architecture has 128 KB per SM [54, 55]. Similarly, AMD CDNA3 data center architecture has 64 KB instruction caches compared to RDNA3 consumer architecture's 32 KB [7].

Memory Configuration: LLM inference decoding is severely memory bandwidth bound and the overall arithmetic intensity is low, so limiting memory bandwidth will significantly reduce AI performance. Gaming applications such as graphics rendering and raytracing on the other hand need to access texture and graphics data stored at different locations in the memory. These irregular accesses are usually latency bound and memory bandwidth utilization is low [76]. Therefore, targeting memory bandwidth becomes the attractive choice to create policies for limiting AI performance. Additionally, data center and consumer GPU devices already have different memory bandwidth configurations - generally data center GPUs use high bandwidth memory while consumer GPUs use lower latency GDDR memory. Limiting device memory capacity reduces the number of model parameters, which correlates with the accuracy and capabilities of models [10]. However, smaller models are able to achieve high accuracy and similar human preferences compared to large models on certain tasks [21].

By scoping regulations to create designs which are inherently limited for AI applications, tailored gaming-focused devices can continue to be sold and reduce overall negative externalities.

6 Related Work

6.1 Computing Sanctions Design

In the United States, computing related sanctions and export controls are administered by the Bureau of Industry and Security (BIS)

under the Department of Commerce. New export controls are implemented via a general federal level rulemaking process [58]. As these regulations are designed, the BIS will solicit guidance from other government departments as well as external stakeholders who may be affected by these regulations. Before regulations are finalized, the BIS will either release a proposed rule or interim final rule which the general public can comment on.

Numerous performance metrics have been previously used to classify computing for export controls. Composite Theoretical Performance (CTP), introduced in 1991, is based on 64 bit FLOPs/sec and includes adjustments for fixed point operations, operation bitwidth scaling, and memory/IO configurations [8]. CTP was replaced in 2006 with Adjusted Peak Performance (APP), which focused on 64 bit FLOPs/sec and weighted vector and non-vector operations differently [73]. APP was later replaced with only theoretical FLOPs/sec [12], and bitwidth scaling was reintroduced with TPP. Performance density has been used in the context of multi-core processors but was designed as an optimization metric using workload performance rather than for device classification using theoretical performance [43].

This work provides a computer architect's perspective on how to optimize designs under existing computing export controls. We demonstrates how combining theoretical based computing classification metrics with informed architectural parameters is more indicative of workload performance. Furthermore, this work proposes economically efficient policies to reduce the negative externality of such regulations which previous works do not mention.

6.2 LLM Inference Hardware

Most previous works regarding improving LLM inference performance have focused on system-level optimizations, including kernel fusion [23], scheduling [4, 59, 80], and parallelism [64]. On the architecture side, LLMCompass covered the compute-bound nature of prefill and memory bandwidth-bound nature of decoding stages and proposed separate throughput and latency oriented designs [82]. This work specifically focuses on the hardware parameters that are affected by current ACRs and shows how these regulations ultimately affect LLM inference performance.

6.3 Domain-Tailored Performance Optimization

In 2021, NVIDIA introduced "Lite Hash Rate" (LHR) gaming GPU designs which had modified firmware to limited the device's Ethereum hash rate without affecting gaming performance [77]. Additionally, NVIDIA introduced CMP Hx line of dedicated to cryptocurrency mining [78]. CMP Hx devices use the same dies as data center and gaming GPUs, but removed display outputs and optimized device voltage and frequency [68, 69].

This work's architecture-first approach for workload specific optimization would encourage creating distinct hardware designs for each workload. However, this does not prevent two different designs from using the same die, as binning dies based on defect locations (*e.g.* cores vs memory PHYs) or purposeful disabling can lead to the same effect. Furthermore, firmware based solutions have seen workarounds to unlock additional performance [74] which architecture-first approaches are less vulnerable to.

7 Conclusion

In conclusion, this work makes the first study on the chip architectural and economic externality implications of the advance computing rules. We demonstrate how the ACR specifications affect chip architecture parameters - TPP essentially limits tensor/matrix core performance and performance density enforces a minimum die area. We conduct a thorough design space exploration of chip architectures for LLM inference under October 2022 and October 2023 ACRs, present quantified performance, die area, and cost tradeoffs of current sanctions, and show how chip architectures can be optimized under these regulations. Under October 2022 specifications, single die designs can still improve TTFT and TBT by 4% and 27% respectively compared to a modeled NVIDIA A100.

From our study, we demonstrate how future policies can be defined to reduce *negative externalities*. Replacing marketing-based device classification with architectural metrics reduces the ambiguity of existing regulations. Combining theoretical performance metrics with select architectural parameters creates better performance indicators. Using TPP and memory bandwidth limits together creates devices with up to 110% slower median TBT latency and up to 42.4x narrower distribution compared to using TPP alone. By guiding regulations towards designs which are inherently limited in performance for workloads-of-interests, manufacturers can continue to improve and sell devices which reduces negative externalities and market distortions.

Current and future events will provide ripe opportunities for computer architects to shape policies - let us leverage our experiences to ensure they are fair, effective, and impactful.

Acknowledgments

We would like to thank Margaret Martonosi for her insights on the second half of this work as well as the USA federal policy making process, Haiyue Ma for our discussions on architecture-first performance indicators, Hengrui Zhang and Rohan Baskar Prabhakar for their help with configuring and extending LLMCompass, as well as Kai Li, the entire Princeton Parallel Group, and our anonymous reviewers for their feedback, suggestions, and encouragement. This material is based upon work supported by a Princeton Andlinger Center Innovation Award, a Princeton SEAS Innovation Award, and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2039656. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Advanced Micro Devices, Inc. 2021. *AMD CDNA™ 2 Architecture.* Technical Report. AMD. https://www.amd.com/content/dam/amd/en/documents/instinct-business-docs/white-papers/amd-cdna2-white-paper.pdf

[2] Advanced Micro Devices, Inc. 2023. *AMD CDNA™ 3 Architecture.* Technical Report. AMD. https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf

[3] Defense Advanced Research Projects Agency. [n. d.]. Circuit Realization at Faster Timescales (CRAFT). https://www.darpa.mil/program/circuit-realization-at-faster-timescales Date Accessed: 7 May 2025.

[4] Amey Agrawal, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, and Ramachandran Ramjee. 2023. SARATHI: Efficient LLM Inference by Piggybacking Decodes with Chunked Prefills. arXiv:2308.16369 [cs.LG]

[5] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query

Transformer Models from Multi-Head Checkpoints. arXiv:2305.13245 [cs.CL] https://arxiv.org/abs/2305.13245

[6] AMD. [n. d.]. AMD FidelityFX Super Resolution. https://www.amd.com/en/products/graphics/technologies/fidelityfx/super-resolution.html Date Accessed: 7 May 2025.

[7] AMD. 2025. Accelerator and GPU hardware specifications. https://rocm.docs.amd.com/en/latest/reference/gpu-arch-specs.html

[8] D. Ames, D. Gibson, and B. Troy. 1991. composite theoretical performance. *SIGMETRICS Perform. Eval. Rev.* 19, 2 (sep 1991), 24–29. doi:10.1145/122564.122565

[9] Wassenaar Arrangement. 2023. Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies. https://www.wassenaar.org/app/uploads/2023/12/List-of-Dual-Use-Goods-and-Technologies-Munitions-List-2023-1.pdf

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in neural information processing systems*, Vol. 33. 1877–1901.

[11] Bureau of Industry and Security. 2023. Public Briefing - Assistant Secretary Thea D. Rozman Kendler 11-06-2023. https://www.youtube.com/watch?v=EvUOwx8mHoo

[12] Bureau of Industry and Security, Department of Commerce. 2007. December 2006 Wassenaar Arrangement Plenary Agreement Implementation: Categories 1, 2, 3, 5 Part I, 6, 7, 8, and 9 of the Commerce Control List; Wassenaar Reporting Requirements; Definitions; and Statement of Understanding on Source Code. Federal Register, 62524-62551 pages. https://www.federalregister.gov/documents/2007/11/05/E7-21247/december-2006-wassenaar-arrangement-plenary-agreement-implementation-categories-1-2-3-5-part-i-6-7-8

[13] Bureau of Industry and Security, Department of Commerce. 2021. Encryption and Export Administration Regulations (EAR). hhttps://www.bis.doc.gov/index.php/policy-guidance/encryption

[14] Bureau of Industry and Security, Department of Commerce. 2022. Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification. Federal Register, 73458-73517 pages. https://www.federalregister.gov/documents/2022/10/13/2022-21658/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor

[15] Bureau of Industry and Security, Department of Commerce. 2023. Commerce Strengthens Restrictions on Advanced Computing Semiconductors, Semiconductor Manufacturing Equipment, and Supercomputing Items to Countries of Concern. Press Release. https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3355-2023-10-17-bis-press-release-acs-and-sme-rules-final-js/file

[16] Bureau of Industry and Security, Department of Commerce. 2023. Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections. Federal Register, 73458-73517 pages. https://www.federalregister.gov/documents/2023/10/25/2023-23055/implementation-of-additional-export-controls-certain-advanced-computing-items-supercomputer-and

[17] Bureau of Industry and Security, Department of Commerce. 2024. Foreign-Produced Direct Product Rule Additions, and Refinements to Controls for Advanced Computing and Semiconductor Manufacturing Items. Federal Register, 96790-96830 pages. https://www.federalregister.gov/documents/2024/12/05/2024-28270/foreign-produced-direct-product-rule-additions-and-refinements-to-controls-for-advanced-computing

[18] Bureau of Industry and Security, Department of Commerce. 2025. Framework for Artificial Intelligence Diffusion. Federal Register, 4544-4584 pages. https://www.federalregister.gov/documents/2025/01/15/2025-00636/framework-for-artificial-intelligence-diffusion

[19] Bureau of Industry and Security, Department of Commerce. 2025. Implementation of Additional Due Diligence Measures for Advanced Computing Integrated Circuits; Amendments and Clarifications; and Extension of Comment Period. Federal Register, 5298-5321 pages. https://www.federalregister.gov/documents/2025/01/16/2025-00711/implementation-of-additional-due-diligence-measures-for-advanced-computing-integrated-circuits

[20] Brian Burke. 2022. NVIDIA Introduces DLSS 3 With Breakthrough AI-Powered Frame Generation for up to 4x Performance. https://nvidianews.nvidia.com/news/nvidia-introduces-dlss-3-with-breakthrough-ai-powered-frame-generation-for-up-to-4x-performance

[21] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 331, 30 pages.

[22] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs.CL]

[23] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *Advances in Neural Information Processing Systems* 35 (2022), 16344–16359.

[24] Hanna Dohmen and Jacob Feldgoise. 2023. A Bigger Yard, A Higher Fence: Understanding BIS's Expanded Controls on Advanced Computing Exports. *Georgetown Center for Security and Emerging Technology* (2023). https://cset.georgetown.edu/article/bis-2023-update-explainer/

[25] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39. http://jmlr.org/papers/v23/21-0998.html

[26] Samantha Fields. 2020. What is market distortion? *Marketplace* (2020). https://www.marketplace.org/2020/11/10/what-is-market-distortion-amazon-eu-antitrust/

[27] Directorate-General for Energy. 2024. Commission adopts EU-wide scheme for rating sustainability of data centres. https://nvidianews.nvidia.com/news/nvidia-introduces-dlss-3-with-breakthrough-ai-powered-frame-generation-for-up-to-4x-performance

[28] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang,

Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[29] Lennart Heim [@ohlennart]. 2023. Here's a visualization of the US export controls on chips. Explicitly designed to target chips with high performance and interconnect... [Image attached][Post]. X. https://x.com/ohlennart/status/1631037103462133761 March 1st, 2023.

[30] International Trade Administration, U.S. Department of Commerce. 2023. SOUTH KOREA SEMICONDUCTORS. https://www.trade.gov/market-intelligence/south-korea-semiconductors

[31] Alun John, Samuel Shen, and Tom Wilson. 2021. China's top regulators ban crypto trading and mining, sending bitcoin tumbling. *Reuters* (2021). https://www.reuters.com/world/china/china-central-bank-vows-crackdown-cryptocurrency-trading-2021-09-24/

[32] Handel Jones. 2015. Semiconductor Industry from 2015 to 2025. *International Business Strategies* (2015). https://www.semi.org/en/semiconductor-industry-2015-2025

[33] Norm Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Clifford Young, Xiang Zhou, Zongwei Zhou, and David A Patterson. 2023. TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings. In *Proceedings of the 50th Annual International Symposium on Computer Architecture* (Orlando, FL, USA) *(ISCA '23)*. Association for Computing Machinery, New York, NY, USA, Article 82, 14 pages. doi:10.1145/3579371.3589350

[34] Saif Khan and Alexander Mann. 2020. AI Chips: What They Are and Why They Matter. (Apr 2020). doi:10.51593/20190014

[35] Saif M. Khan, Alexander Mann, and Dahlia Peterson. 2021. The Semiconductor Supply Chain: Assessing National Competitiveness. (Jan 2021). doi:10.51593/20190016

[36] Lawrence Livermore National Laboratory. [n. d.]. El Capitan: Preparing for NNSA's first exascale machine. https://asc.llnl.gov/exascale/el-capitan Date Accessed: 7 May 2025.

[37] Jane Lee. 2022. Exclusive: Nvidia offers new advanced chip for China that meets U.S. export controls. *Reuters* (2022). https://www.reuters.com/technology/exclusive-nvidia-offers-new-advanced-chip-china-that-meets-us-export-controls-2022-11-08/

[38] Jane Lanhee Lee and Mackenzie Hawkins. 2024. AMD Hits US Roadblock in Selling AI Chip Tailored for China. *Bloomberg* (2024). https://www.bloomberg.com/news/articles/2024-03-05/amd-hits-us-roadblock-in-selling-ai-chip-tailored-for-china

[39] Yen Nee Lee. 2019. The Japan-South Korea dispute could push up the price of your next smartphone. *CNBC* (2019). https://www.cnbc.com/2019/07/23/japan-south-korea-dispute-impact-on-semiconductor-supply-chain-prices.html

[40] List of AMD graphics processing units [n. d.]. https://en.wikipedia.org/wiki/List_of_AMD_graphics_processing_units Date Accessed: 23 April 2025.

[41] List of Nvidia graphics processing units [n. d.]. https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units Date Accessed: 23 April 2025.

[42] Lufei Liu, Mohammadreza Saed, Yuan Hsi Chou, Davit Grigoryan, Tyler Nowicki, and Tor M Aamodt. 2023. LumiBench: A Benchmark Suite for Hardware Ray Tracing. In *2023 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 1–14.

[43] Pejman Lotfi-Kamran, Boris Grot, Michael Ferdman, Stavros Volos, Onur Kocberber, Javier Picorel, Almutaz Adileh, Djordje Jevdjic, Sachin Idgunji, Emre Ozer, and Babak Falsafi. 2012. Scale-out processors. In *Proceedings of the 39th Annual International Symposium on Computer Architecture* (Portland, Oregon) *(ISCA '12)*. IEEE Computer Society, USA, 500–511.

[44] N. Gregory Mankiw. 2011. *Principles of Microeconomics* (6 ed.). Cengage Learning, Stamford, CT.

[45] Yelin Mo and Fanny Potkin. 2024. Nvidia to launch China-focused AI chip in Q2 2024 - sources. *Reuters* (2024). https://www.reuters.com/technology/nvidia-launch-china-focused-ai-chip-q2-2024-sources-2024-01-08/

[46] Stephen Nellis and Jane Lee. 2023. Nvidia tweaks flagship H100 chip for export to China as H800. *Reuters* (2023). https://www.reuters.com/technology/nvidia-tweaks-flagship-h100-chip-export-china-h800-2023-03-21/

[47] August Ning, Georgios Tziantzioulis, and David Wentzlaff. 2023. Supply Chain Aware Computer Architecture. In *Proceedings of the 50th Annual International Symposium on Computer Architecture* (Orlando, FL, USA) *(ISCA '23)*. Association for Computing Machinery, New York, NY, USA, Article 17, 15 pages. doi:10.1145/3579371.3589052

[48] August Tianyin Ning. 2025. *Computer Architecture Under Economic Constraints*. PhD Thesis. Princeton University, Princeton, NJ.

[49] NVIDIA. 2022. NVIDIA A30 TENSOR CORE GPU. https://www.nvidia.com/content/dam/en-zz/Solutions/data-center/products/a30-gpu/pdf/a30-datasheet.pdf

[50] NVIDIA. 2023. GeForce RTX 4090 D. https://www.nvidia.cn/geforce/graphics-cards/40-series/rtx-4090-d/

[51] NVIDIA. 2023. NVIDIA Driver License Agreement. https://www.nvidia.com/en-us/drivers/geforce-license/

[52] NVIDIA A800 40GB Active 2023. https://resources.nvidia.com/en-us-briefcase-for-datasheets/proviz-a800-40gb-dat

[53] NVIDIA Corporation. 2020. *NVIDIA A100 Tensor Core GPU Architecture*. Technical Report. NVIDIA. https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

[54] NVIDIA Corporation. 2022. *NVIDIA H100 Tensor Core GPU Architecture*. Technical Report. NVIDIA. https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper

[55] NVIDIA Corporation. 2023. *NVIDIA Ada GPU Architecture*. Technical Report. NVIDIA. https://images.nvidia.com/aem-dam/Solutions/Data-Center/l4/nvidia-ada-gpu-architecture-whitepaper-v2.1.pdf

[56] NVIDIA H800 Tensor Core GPU 2023. https://resources.nvidia.com/en-us-briefcase-for-datasheets/proviz-a800-40gb-dat

[57] National Oceanic and Atmospheric Administration. 2025. High Performance Computing and Communications. https://www.noaa.gov/information-technology/hpcc
[58] The Office of the Federal Register. 2013. A Guide to the Rulemaking Process. https://www.federalregister.gov/uploads/2013/09/The-Rulemaking-Process.pdf
[59] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. Splitwise: Efficient Generative LLM Inference Using Phase Splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*. 118–132. doi:10.1109/ISCA59077.2024.00019
[60] Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. 2024. Computing Power and the Governance of Artificial Intelligence. arXiv:2402.08797 [cs.CY] https://arxiv.org/abs/2402.08797
[61] Noam Shazeer. 2020. GLU Variants Improve Transformer. arXiv:2002.05202 [cs.LG] https://arxiv.org/abs/2002.05202
[62] David Shepardson. 2023. US in talks with Nvidia about AI chip sales to China - Raimondo. *Reuters* (2023). https://www.reuters.com/technology/us-talks-with-nvidia-about-ai-chip-sales-china-raimondo-2023-12-11/
[63] Anton Shilov. 2023. Understanding the Big Spend on Advanced Packaging Facilities. *EE Times* (2023). https://www.eetimes.com/understanding-the-big-spend-on-advanced-packaging-facilities/
[64] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* (2019).
[65] Gyung-hwa Song. 2019. Semiconductor industry insiders calls for less dependence on Japanese imports. *The Hankyoreh* (2019). https://english.hani.co.kr/arti/english_edition/e_international/900366.html
[66] Swiss Datacenter Efficiency Association [n. d.]. https://www.sdea.ch/
[67] TechPowerUp. [n. d.]. GPU Specs Database. https://www.techpowerup.com/gpu-specs/ Date Accessed: 23 April 2025.
[68] TechPowerUp. 2020. https://www.techpowerup.com/gpu-specs/nvidia-ga100.g931
[69] TechPowerUp. 2020. https://www.techpowerup.com/gpu-specs/nvidia-ga102.g930
[70] TechPowerUp. 2023. https://www.techpowerup.com/gpu-specs/nvidia-ad104.g1013
[71] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] https://arxiv.org/abs/2302.13971
[72] Kristin Uchiyama. 2022. NVIDIA Announces Hopper Architecture, the Next Generation of Accelerated Computing. https://nvidianews.nvidia.com/news/nvidia-announces-hopper-architecture-the-next-generation-of-accelerated-computing
[73] Information Systems Technical Advisory Committee US Dept. of Commerce, BIS. 2006. A PRACTITIONER'S GUIDE TO ADJUSTED PEAK PERFORMANCE. https://www.bis.doc.gov/index.php/documents/product-guidance/865-practioner-s-guide-to-adjusted-peak-performance/file
[74] NBMiner v39.0. 2021. https://github.com/NebuTech/NBMiner/releases/tag/v39.0
[75] VideoCardz. 2024. NVIDIA to launch HGX H20, L20 and L2 GPUs for China. *VideoCardz* (2024). https://videocardz.com/newz/nvidia-to-launch-hgx-h20-l20-and-l2-gpus-for-china
[76] Peng Wang and Zhibin Yu. 2023. RayBench: An Advanced NVIDIA-Centric GPU Rendering Benchmark Suite for Optimal Performance Analysis. *Electronics* 12, 19 (2023), 4124.
[77] Matt Wuebbling. 2021. A Further Step to Getting GeForce Cards into the Hands of Gamers. *NVIDIA* (2021). https://blogs.nvidia.com/blog/lhr/
[78] Matt Wuebbling. 2021. GeForce Is Made for Gaming, CMP Is Made to Mine. *NVIDIA* (2021). https://blogs.nvidia.com/blog/geforce-cmp/
[79] Falan Yinug. 2021. Chipmakers Are Ramping Up Production to Address Semiconductor Shortage. Here's Why that Takes Time. https://www.semiconductors.org/chipmakers-are-ramping-up-production-to-address-semiconductor-shortage-heres-why-that-takes-time/
[80] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. 2022. Orca: A Distributed Serving System for Transformer-Based Generative Models. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*. USENIX Association, Carlsbad, CA, 521–538. https://www.usenix.org/conference/osdi22/presentation/yu
[81] Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, Yan Yan, Beidi Chen, Guangyu Sun, and Kurt Keutzer. 2024. LLM Inference Unveiled: Survey and Roofline Model Insights. arXiv:2402.16363 [cs.CL] https://arxiv.org/abs/2402.16363
[82] Hengrui Zhang, August Ning, Rohan Baskar Prabhakar, and David Wentzlaff. 2024. LLMCompass: Enabling Efficient Hardware Design for Large Language Model Inference. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*. 1080–1096. doi:10.1109/ISCA59077.2024.00082