Supply Chain Aware Computer Architecture

August Ning
Princeton University
Princeton, NJ, USA
aning@princeton.edu

Georgios Tziantzioulis
Princeton University
Princeton, NJ, USA
georgios.tziantzioulis@pm.me

David Wentzlaff
Princeton University
Princeton, NJ, USA
wentzlaf@princeton.edu

ABSTRACT

Progressively and increasingly, our society has become more and more dependent on semiconductors and semiconductor-enabled products and services. The importance of chips and their supply chains has been highlighted during the 2020-present chip shortage caused by manufacturing disruptions and increased demand due to the COVID-19 pandemic. However, semiconductor supply chains are inherently vulnerable to disruptions and chip crises can easily recur in the future.

We present the first work that elevates supply chain conditions to be a first-class design constraint for future computer architectures. We characterize and model the chip creation process from standard tapeout to packaging to provide a framework for architects to quickly assess the time-to-market of their chips depending on their architecture and the current market conditions. In addition, we propose a novel metric, the Chip Agility Score (***CAS***) - a way to quantify a chip architecture's resilience against production-side supply changes.

We utilize our proposed time-to-market model, ***CAS***, and chip design/manufacturing economic models to evaluate prominent architectures in the context of current and speculative supply chain changes. We find that using an older process node to re-release chips can decrease time-to-market by 73%-116% compared to using the most advanced processes. Also, mixed-process chiplet architectures can be 24%-51% more agile compared to equivalent single-process chiplet and monolithic designs respectively. Guided by our framework, we present an architectural design methodology that minimizes time-to-market and chip creation costs while maximizing agility for mass-produced legacy node chips.

Our modeling framework and data sets are open-sourced to advance supply chain aware computer architecture research. https://github.com/PrincetonUniversity/ttm-cas

CCS CONCEPTS

• **Hardware → Economics of chip design and manufacturing**; **VLSI design manufacturing considerations**; • **Computer systems organization → Architectures**.

KEYWORDS

semiconductor supply chain, chip shortage, modeling, economics

ISCA '23, June 17–21, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0095-8/23/06.
https://doi.org/10.1145/3579371.3589052

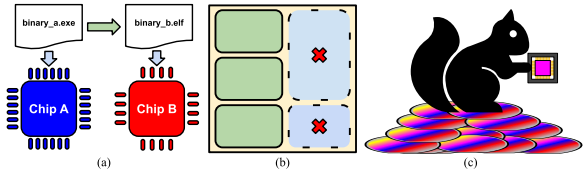

Figure 1: In the face of chip shortages, companies have (a) rewritten software to use available chips [30]; (b) shipped products with missing features [23]; (c) hoarded chips, which has exacerbated shortages [69]

ACM Reference Format:
August Ning, Georgios Tziantzioulis, and David Wentzlaff. 2023. Supply Chain Aware Computer Architecture. In *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23), June 17–21, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3579371.3589052

1 INTRODUCTION

Over the past century, semiconductors have increasingly been adopted in our daily lives in both critical and non critical systems; the insatiable demand continues to grow. In 2021, the global semiconductor industry sold 1.15 trillion units totaling $555.9 billion USD in revenue, and sales are projected to cross $1 trillion USD by 2030 [11, 17].

The importance of chips and chip manufacturing supply chain disruptions have been highlighted by multiple events over the past years: the 2011 Thailand floods halted global hard drive manufacturing capacity - a single factory in Bang Pa-in Thailand produced 25% of the world's hard drive head sliders; 2019 to present COVID-19 pandemic has led to a global chip shortage affecting various sectors; the 2022 conflict in Eastern Europe has led to sanctions and material shortages crucial to semiconductor manufacturing [6, 7, 37, 81]. The consequences of the recent shortages have been widely discussed spanning the fields of semiconductor engineering and manufacturing, supply chain management, economics, and policy [32, 47, 67, 82, 97, 99, 126]. Chip designers, manufacturers, and users have scrambled to adapt to these changes, as featured in Figure 1. In late 2022 through 2023, looming economic downturn has led to decreased chip demand which may again affect foundry capacity and chip supply chains [87].

As computer architecture approaches the limits of performance and efficiency, architects must consider the time-to-market and supply chain vulnerability of their designs. This is the **first work to propose elevating supply chain conditions as first-class design constraints in computer architecture research**. Also, this is the first work to introduce a time-to-market model and supply chain agility metric for computer architecture. By characterizing the chip creation process from tapeout through packaging, we provide **a public and open-sourced modeling framework** for

designers and manufacturers to quickly and quantitatively assess the **time-to-market** of their chips depending on its architecture and current market conditions.

Using this framework, we propose a novel metric, **the Chip Agility Score (*CAS*) to evaluate a chip architecture's agility against production side supply chain changes.** *CAS* is based on the rate of change of time-to-market relative to foundry production capacity; designs with higher agility are more resilient to production-side supply chain changes.

We evaluate our time-to-market model, *CAS*, and an updated chip creation cost model [56] in the context of estimated current market conditions and make the following observations: ① time-to-market optimized designs are architecturally different compared to optimizing for performance and cost; ② lower time-to-market does not necessarily correlate with higher *CAS* or higher chip creation costs; ③ packaging silicon dies from multiple process nodes may be more agile than single-process node chiplets depending on supply chain disruption severity.

Guided by time-to-market, *CAS*, and cost modeling, we evaluate a chip design methodology of designing and manufacturing the same architecture concurrently on multiple process nodes. Our framework identifies optimal process combinations and production splits that maximize *CAS* while minimizing time-to-market and chip creation costs. For our study, the fastest multi-process split is 47% more agile than the fastest single process and is 8% faster-to-market compared to the cheapest process while only increasing costs by 1.6%.

This work makes the following contributions:

- Detailed overview of the currently inflexible and vulnerable nature of chip creation and semiconductor manufacturing supply chains.
- The first work to propose using supply chains and time-to-market metrics as first-class design constraints in computer architecture research.
- Proposal of an open-source chip creation time-to-market modeling framework and Chip Agility Score (*CAS*) to allow architects for the first time to quantitatively evaluate an architecture's time-to-market and agility in the presence of supply chain changes.
- Evaluation of time-to-market, *CAS*, and chip creation costs on prominent architectures in the context of current and future production supply chain changes based on public, published, and derived parameters.

2 BACKGROUND AND MOTIVATION

In this section, we provide background on the chip creation process, outline the difficulties of chip creation, and analyze fabrication supply chain vulnerabilities in current and future contexts. We then motivate how a computer architect's design choices can affect the time-to-market of their chips and the importance of modeling the chip creation process.

2.1 Chip Creation Process

The chip creation process is an intricate procedure. Figure 2 presents an abstracted and generalized overview of the process and captures its core steps.

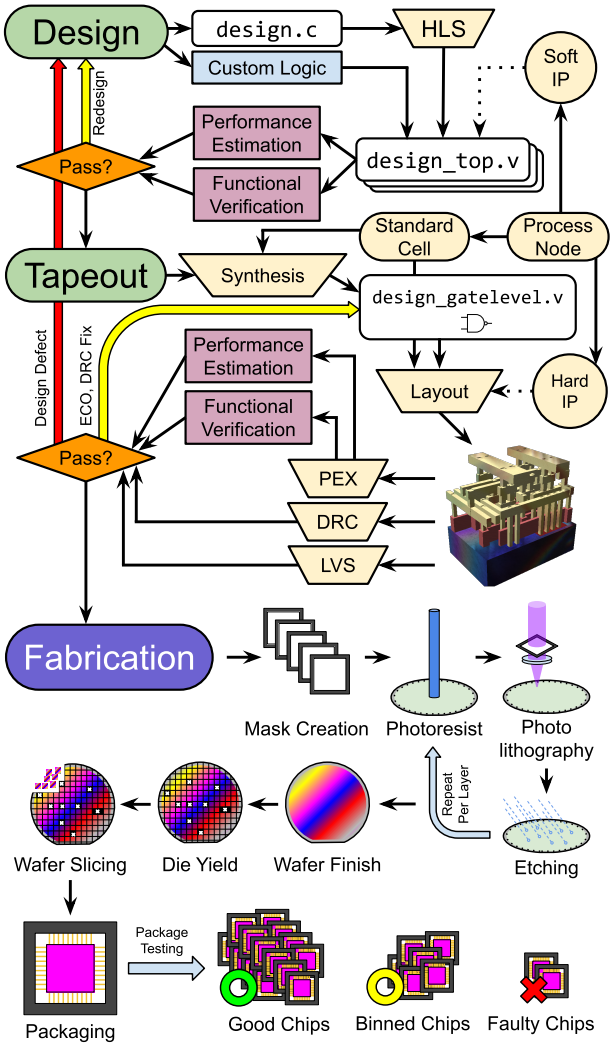

Figure 2: Chip Creation Process Summary

The process begins with design, implementation, and verification where computer architects must devise a chip that implements the desired functionality and meets the specification and underlying technology constraints: performance, thermal design power, die area, clock frequency, *etc.* The implementation is most commonly done using register-transfer level (RTL) languages and high-level synthesis (HLS) tools. Once the implementation is completed, an initial round of verification is performed with the RTL design to ensure that preliminary functionality and timing specifications are met which may lead to additional modifications. Based on verification results, additional modifications are performed as needed to meet the target specifications.

Next, the first step of the tapeout phase is to synthesize the design from RTL to a logic gate-level description via synthesis EDA tools. Subsequently, the gate-level description is used by place-and-route tools which will layout the logic gates (implemented using standard cell libraries) and route metal layers to connect the logic gates. The place-and-route tools also integrate any hand-designed, "full

custom" circuits which are designed and laid out at the transistor granularity.

The gate-level design is then checked to ensure the foundry's and process node's design rules are satisfied (DRC) and the layout is checked against the logical schematic of the gate-level design to ensure that metal layers were routed correctly (LVS). EDA tools then analyze the layout, extract a parasitic netlist, and engineers perform gate-level design verification. When the design has been verified to meet functionality and timing requirements, it can be signed-off and sent to the foundry for fabrication.

After receiving the layout, the foundry creates the photolithography masks that are used to pattern the design, etch the silicon, and deposit the metal layers onto the wafer. The silicon wafer is then processed, cut into dies, tested, and packaged. These last steps can be done by the foundry or at outsourced semiconductor assembly and test (OSAT) firms. OSATs may test the dies and/or the final packages and only package/return the chips that meet the customer's specifications. Customers may choose to separate chips by their performance characteristics or defects, commonly known as "binning". The chip creation process is completed once the dies are packaged and shipped back to the customer.

2.2 Chip Creation Challenges

The chip creation process is expensive, complex, and time consuming. Starting with the tapeout phase, running EDA tools takes significant computation and human effort [92]. Additionally, any design rule violations, timing violations, verification errors, and critical design defects may require the tapeout phase to be rerun or incur costly redesign. Design rule violations must be fixed by place-and-route tools with new layout constraints or hand fixed by the engineer. Static timing violations may be fixed with an engineering change order (ECO), but then design rule checks and layout-schematic verification must be rerun. Layout functional and dynamic timing verification requires engineers to create and run test cases and identify any design defects before the design is sent to the foundry.

The complexity of the tapeout process is correlated with the target process node used for the fabrication. With smaller transistors, foundries require more complicated design rules which the EDA tools must take into consideration during place-and-route [108]. Higher transistor and wire density at advanced process nodes mean EDA tools need to check additional neighboring transistors and wires for violations. The increase in verification complexity means tapeout time and costs grow exponentially with more advanced process nodes [63, 92].

Chip fabrication is highly specialized and requires expensive capital investments, so only a handful of firms are capable of creating chips for customers. In 2022, the top five semiconductor foundries accounted for over 90% of the global market share by revenue, with top player Taiwan Semiconductor Manufacturing Company (TSMC) controlling a majority of the market [60]. Additionally, production at bleeding-edge process nodes ("sub 10nm") is only known to be achieved at scale by three firms [33, 45, 119]. Due to the small number of fabs and the ever-growing demand for semiconductors, chip designers need to plan far in advance to secure foundry capacity in the future or face long lead times for their chips to be produced.

Advanced process nodes have higher transistor density meaning defects of the same area on a wafer can affect more transistors than at less dense, legacy process nodes [114]. Moreover, advanced process nodes may operate a less mature flow and it takes time for foundries to perfect the manufacturing process - wafer yield (the percentage of functional dies from a single wafer) is expected to increase the longer the process node is in production [27]. The lower yield rate of advanced process nodes can increase the number of chips a customer needs to order from the foundry which in turn increases lead times for all chips to be fabricated.

This complicated process results in high design and manufacturing costs. In order for chip designers to profit, products must meet time-to-market requirements to maximize revenue [89].

2.3 Current and Future Supply Chain Challenges

The 2020-present chip shortage has been prominently featured in news coverage and dominated contemporary supply chain concerns. Stemming from the bounce-back in consumer demand in 2021, semiconductor foundries have been inundated with orders and lead times had almost doubled [58], global foundry capacity was fully booked [68, 71], and automotive and consumer electronic sales have been severely affected [80, 95, 96, 124].

Although the current chip shortage is considered temporary, the chip fabrication supply chain is inherently inflexible and susceptible to future supply chain crises. In the field of supply chain management, a supply chain can be evaluated on its *resilience* to supply chain disruptions depending on its *vulnerabilities* to disruptions and its *capabilities* to adapt to them; the chip fabrication supply chain faces many vulnerability factors and has few capabilities to react to them [88].

Firstly, the specialized fabrication process makes it difficult to source materials and equipment from alternative suppliers, especially for silicon wafers and extreme ultraviolet lithography (EUV) lithography machines [22, 116]. TSMC sources raw silicon wafers from five suppliers who control over 90% of global silicon wafer supply [116]. The most advanced process nodes require EUV machines which are only produced by ASML [22]. The importance of semiconductors within a national security context combined with recent geopolitical tensions has led to substantial embargos on key components required for chip creation [40, 77, 98, 112].

In addition, foundries are capital-intense and take a long time to build and ramp up production - new foundries take three to four years of construction until production can begin [39, 46, 68]. The complex process also prevents fabs from quickly starting production after shutdowns as seen in 2021 with Texas snow storms and Reneseas' foundry fire [25, 31, 93]. Climate change has worsened droughts in Taiwan and in the southwestern United States, where major advanced process node foundries are located which strains production capacity [20, 102, 127]. Recent and historical flooding in Southeast Asia has affected semiconductor manufacturing and packaging supply chains [53, 91, 94].

The foundries also lack transparency on their internal supply chains and aggregate customer demand [70]. Architects have to choose a design's foundry and process node during the design phase, as standard cell libraries and physical IP are specific to both [113].

Table 1: Chip Creation Process Model Parameters

Parameter	Explanation
N_{TT}	Number of Total Transistors
N_{UT}	Number of Unique/Unverified Transistors
$E_{tapeout}$	Tapeout Engineering Effort
N_W	Number of Wafers
μ_W	Wafer Production Rate of the Foundry
L_{fab}	Foundry Fabrication Latency
n	Number of Final Chips
Y	Die Yield
A_{die}	Die Area
$N_{die,package}$	Number of Dies per Package
L_{TAP}	Testing, Assembly, and Packaging Latency
$E_{testing}$	Testing Engineering Effort
$E_{packaging}$	Packaging Engineering Effort

This makes it complicated for customers to choose the foundry that is able to produce their chips the fastest during a supply chain crisis.

The future demand on foundries will be for more chips from an increasingly diverse set of firms - an increasing number of organizations that have not traditionally built chips such as consumer electronics giants [9, 41, 44, 65, 100], bespoke hardware startups [2, 73, 104], and non-traditional fabless companies (most notably in the automotive space) [3, 14, 74, 125] are designing their own chips and require foundries to fabricate them. With more customers, it becomes difficult to predict aggregate consumer demand and foundries become more vulnerable to demand shocks.

2.4 Motivation

Semiconductors are paramount in global commerce and the recent chip shortage has highlighted the importance of their supply chains. The prominent role of semiconductors in both the economy and security has raised securing chips to a top government concern - the United States, China, South Korea, and the European Union all recently passed groundbreaking legislation for funding billions of USD into semiconductor manufacturing and research [57, 59, 99, 123, 126]. However, even with additional investments, it takes significant time to bring up new semiconductor foundries and the inflexible chip supply chain remains vulnerable to future shortages.

During a chip supply chain crisis, firms are bottle-necked by the foundries' production capacity [81]. Traditionally, architects generally design chips within functional, performance, power, and cost constraints. Equally important, architects need to now consider how their RTL level design, process node choice, and packaging configuration affect the time-to-market of their chips. The tradeoffs made during the design phase dictate how easily a design can be modified to accommodate changing market conditions and therefore resilience against future chip shortages.

The semiconductor design and manufacturing industry is incredibly secretive and architects are hampered from making informed decisions. It is unclear what internal modeling may already be implemented - no previous time-to-market models have been publicized that account for chip architecture. By creating an open-sourced chip creation time-to-market model, both designers and manufacturers can quickly evaluate chip architectures before completing the design process. Given the importance of the global chip shortage, this work hopes to highlight the need for more public reporting to support research on alleviating current and preventing future chip crises.

3 CHIP CREATION PROCESS MODEL

In this section, we introduce our chip creation model and demonstrate how the computer architect's design tradeoffs and current market conditions affect time-to-market. A summary of model parameters is shown in Table 1.

At a high level, the time-to-market (TTM) of a given design for the current state of the supply chain can be captured through modeling the required time for design and implementation, tapeout, fabrication, and packaging of the design as shown in Eq. 1.

$$TTM = T_{design+implementation} + T_{tapeout} + T_{fabrication} + T_{package} \quad (1)$$

3.1 Design and Implementation Phase

The design and implementation phase is entirely dependent on the engineers and independent of supply chain conditions. Additionally, it is difficult to draw a relationship between a chip's architecture and the design and implementation time: a homogeneous multicore processor may only require the design time for the single core but may be repeated many times to create a large chip that is prone to low yield; an ASIC optimized for energy efficiency may have few transistors but require many design iterations to meet their design goals. Due to the specific nature of the design and implementation phase for each chip architecture, for the purposes of this work we assume design and implementation time can be modeled through a per design constant.

3.2 Tapeout Phase

The amount of time spent in the tapeout phase is related to the size of the design (*e.g.* the number of unqiue and unverified transistors in the gate-level design) and the required time for the EDA/CAD tools to process the design (*e.g.* time it takes to synthesize RTL, run place-and-route, *etc.*)

During the tapeout phase, transistors from the gate-level design needs to be laid out and pass design rule checks, but it may not be necessary to layout and verify every transistor. Chips are built in block level increments that can be reused across different parts of the design and a single block only needs to complete the tapeout phase once. For example, a multicore processor with identical cores can be taped-out by first performing a tape-in of a single core and then using the verified core block for the tapeout of the full processor. For the top level, only the interconnect logic between the cores has to be completed in the tapeout phase. Similarly, using gate-level soft and/or IP cores reduces tapeout time as the vendors have already verified them. To take this into account, our model only considers the unique and unverified transistors that need to complete the tapeout phase.

The number of unique transistors can be quickly estimated from previous designs or from the amount of unique RTL code. An accurate transistor count can be achieved by running the synthesis tool on the RTL design and counting the number of standard cells used. This may be feasible as the synthesis tool generally runs faster than the rest of the tapeout phase's tools [92].

The tools' tapeout phase run times are a function of the chip's target process node. As previously discussed, a foundry's required design rules become more complex and numerous as the process

nodes shrink. Additionally, the final chip design may contain dies that are created at different process nodes, so total tapeout time must the sum of tapeout times across all process nodes p_i in the design d. Therefore, the tapeout time in engineering-hours for a certain chip design $T_{tapeout}$ can be modeled with Eq. 2:

$$T_{tapeout} = \sum_{p_i \in d} N_{UT}(d, p_i) * E_{tapeout}(p_i) \tag{2}$$

Where $N_{UT}(d, p_i)$ refers to the number of unique and unverified transistors for a chip design d that are taped out in process node p_i, and $E_{tapeout}(p_i)$ refers to the tapeout engineering effort of process node p_i. Note that $T_{tapeout}$ is in terms of engineering-hours and the total time it takes to complete the tapeout phase depends on the chip's design hierarchy, the blocks that can be taped out in parallel, and the number of tapeout engineers.

3.3 Fabrication Phase

After the tapeout phase, architects have little control over the rest of the chip creation process and these downstream phases are the ones most crucial during chip shortages. The fabrication phase accounts for the time it takes to fab the design and can be split into queuing and production stages. The queuing stage is the waiting time before production begins and the production stage is when the customer's design is fabricated. The packaging phase is a synchronization point in the chip creation process, as all types of dies that are needed in the final chip have to arrive before packaging can begin. Therefore the overall fabrication time is dependent on the die that takes the longest during the fabrication phase and is shown in Eq. 3.

$$T_{fab} = \max_{p_i \in d} \left(T_{fab,queue}(p_i) + T_{fab,prod}(d, n, p_i) \right) \tag{3}$$

The queuing and production times are dependent on the foundry's demand for chips from customers and the production rates at specific process nodes. With multiple firms and industries all making orders to the few foundries, it is difficult to collect the full market intelligence to model the foundries' demand. Therefore, the queuing time is derived from the lead times quoted from the foundries for a specific process node. The lead time can also be represented with Eq. 4:

$$T_{fab,queue} = \frac{N_{W,ahead}(c, p)}{\mu_W(c, p)} \tag{4}$$

where $N_{W,ahead}(c, p)$ refers to the number of wafers ahead of the current design at process node p at current market conditions c and $\mu_W(c, p)$ refers to the foundry's wafer production rate at process node p at current market conditions c (usually quoted in kilo-wafers per month).

Semiconductor fabrication requires a complex internal assembly line and the processing time of a single wafer lot ($\sim$ 25 wafers) from start to finish can range from 12 to 20 weeks [16, 62, 128]. We refer to this fabrication time as the process node's foundry latency. Assuming an efficient and pipelined assembly line where a new wafer lot can begin production once another lot finishes, the production time can be modeled with Eq. 5:

$$T_{fab,prod} = \frac{N_W(d, n, p)}{\mu_W(c, p)} + L_{fab}(p) \tag{5}$$

with $N_W(d, n, p)$ representing the number of wafers required for producing n final chips of design d at process node p and $L_{fab}(p)$ representing the foundry's latency for process node p. This phase accounts for the expected die yield of the design, as firms must order enough wafers to create the desired number of final chips in expectation of fabrication defects; these additional wafers will increase $N_W(d, n, p)$ and in turn $T_{fab,prod}$.

3.4 Packaging Phase

The packaging phase is the time it takes the packaging house to test and assemble the dies into the final package to complete the chip. To account for the die yield rate, this model uses a negative binomial yield distribution to account for die yield rate $Y(A_{die}(d, p), p)$ shown in Eq. 6 [26]:

$$Y(A_{die}(d, p), p) = \left(1 + \frac{A_{die}(d, p) * D_0(p)}{\alpha} \right)^{-\alpha} \tag{6}$$

where $A_{die}(d, p)$ is die area for the design d at process node p, $D_0(p)$ the defect density of the process node p, and α the cluster parameter.

While it is possible for a firm to test and package their finished dies with a separate OSAT firm, many foundries (including all of the top five foundries) either directly package or partner with packaging firms to provide basic to advanced packaging services [34, 38, 106, 117, 121]. In our model, we assume that the design is packaged with the foundry/foundry partner and fabricated wafers can immediately begin packaging.

Similar to the foundries, the testing, assembly, and packaging firms also have internal assembly lines and queues with an inherent baseline latency for the die/package to complete the phase [128]. Related to the chip design itself, a higher overall transistor count corresponds to additional testing time, as all of the transistors on a die must be tested and only the qualified dies are packaged. Related to packaging, larger dies require additional packaging time as they generally have more pins which require more labor and machine time [76]. Similarly, chips that have more dies per package (such as multichip-modules/"chiplets") incur additional packaging alignment effort and labor time [64, 76].

With these considerations in mind, the packaging phase time $T_{package}$ can be modeled with Eq. 7:

$$\begin{aligned} T_{package} = {} & L_{TAP} \\ & + \left(\frac{n}{Y(A_{die}(d, p), p)} \right) * N_{TT,die}(d) * E_{testing}(p) \\ & + n * N_{die,package}(d) * A_{die}(d, p) * E_{package}(p) \end{aligned} \tag{7}$$

where L_{TAP} is the baseline testing, assembly, and packaging latency, n the number of final chips, $N_{TT,die}(d)$ the number of total transistors per die for design d, $N_{die,package}(d)$ the number of dies packaged per final chip, $A_{die}(d, p)$ the area of the die for design d at process node p, $E_{testing}(p)$, $E_{package}(p)$ the engineering effort for testing and packaging respectively. Due to die yield loss, more than n dies will need to be tested to find enough qualified dies to package.

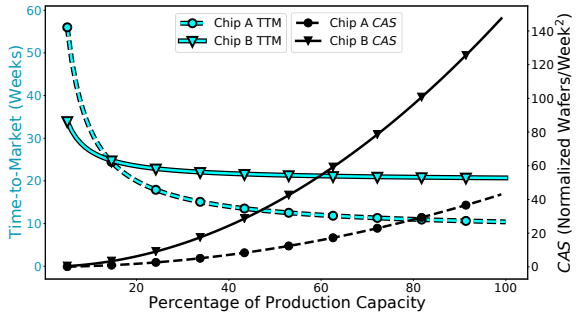

Figure 3: Time-to-Market and *CAS* of Chip A and Chip B

4 CHIP AGILITY SCORE

The chip creation model is focused on the time-to-market of a chip depending on its architecture and current supply chain conditions. It aims to provide insight to chip architects and supply chain analysts on how their design decisions affect how quickly their chips can complete the creation process. However, only evaluating this model in the current supply chain context does not provide insight into a chip's architecture's inherent resilience, or sensitivity, to supply chain disruptions.

Production capacity changes require nuanced study, as decreases in production rates also increase $T_{fab,queue}$ and $T_{fab,prod}$. To evaluate how a specific chip architecture fares in the face of production side supply chain changes, we introduce the **Chip Agility Score** (***CAS***). Consider the time-to-market curves for chips Chip A and Chip B which are fabricated for the same number of final chips as represented by the cyan lines in Figure 3 (left axis). As production capacity decreases, Chip A's time-to-market increases at a faster rate compared to Chip B's which means Chip A's architecture is more sensitive to the foundry's wafer production rate. This rate of change we define as $\frac{\partial}{\partial \mu_W} TTM$. Accordingly, because Chip B has a lower time-to-market rate of change, this architecture is considered more agile against supply chain interruptions even though it has a higher time-to-market at max production rate.

Time-to-market generally increases as production rate decreases, meaning $\frac{\partial}{\partial \mu_W} TTM$ is negative, so we only consider its absolute value. Additionally, agility is dependent on all process nodes p_i that the design d uses, as supply chain disruptions on any of the process nodes can delay the packaging phase for the final chip. Finally, in order to reward lower $\left|\frac{\partial}{\partial \mu_W} TTM\right|$, we take the inverse of sum of the rate of change(s) to make a higher *CAS* reflect a more agile architecture. *CAS* is defined in Eq. 8 and the *CAS* for Chip A and Chip B are represented by the black lines in Figure 3 (right axis).

$$CAS = \left(\sum_{p_i \in d} \left| \frac{\partial}{\partial \mu_W(p_i)} TTM(c, d, n, p_i) \right| \right)^{-1} \tag{8}$$

$TTM(c, d, n, p)$ is defined as the time-to-market of a chip design d for n number of final chips at the used process node p in market conditions c and $\mu_W(p)$ the wafer production rate of process node p. *CAS* is measured in wafers per week squared.

In essence, *CAS* captures and quantifies how agile a design is to production-side changes. **A higher *CAS* indicates that a chip's architecture is less bottle-necked by the chip creation process to produce the desired final number of chips compared to architectures with lower *CAS*.** If the supply chain changes dramatically but the time-to-market of a design does not significantly change, it has a higher *CAS*.

CAS needs to be evaluated within the context of the final number of chips as both $T_{fabrication}$ and $T_{package}$ have assembly line latency times that do not scale with foundry production rates. When producing a few final chips, the design's time-to-market is dominated by these latencies. *CAS* also needs to take into account current market conditions, as high consumer demand and/or low wafer production rates at certain process nodes drastically affect time-to-market. *CAS* does not take into account $T_{design+implement.}$ and $T_{tapeout}$ as they are upstream and independent of the production rate in the chip creation process.

5 METHODOLOGY AND VALIDATION

To best model the current state of the chip creation process, we derive the values for our model from public and published work. [1] All public information, publications, and methodology used to derive model parameters are disclosed throughout this work using citations. In addition, our open-source modeling framework allows users to easily plug in their values and availability for their particular chip designs.

Tapeout engineering effort $E_{tapeout}(p)$ and packaging effort $E_{package}(p)$ are derived from the verification costs and physical costs respectively of each process node from [48, 49, 63], as well as our own experience with the chip creation process; we curve fit to an exponential regression. Testing effort $E_{testing}(p)$ is derived from validation costs from [63] and the projected minimum test data volume from [1] and fit a linear regression.

A foundry's wafer production rate $\mu_W(c, p)$ is derived from earnings reports from publicly traded foundries [118] and the estimated wafer cost per process node from [54]. Wafer production rates per process node are shown in Table 2. We assume the foundry knows enough about its internal supply chain to route orders such that their production rate can be used in the aggregate for our model and that the production rate is used to produce the design's wafers.

Defect densities $D_0(p)$ are based on numbers reported from [27, 111]: D_0 is set low for legacy nodes to represent a mature manufacturing process and increase starting from 20nm. The cluster parameter is set to $\alpha = 3$ to model average defect clustering [111]. Foundry and packaging latency $L_{fab}(p)$, L_{TAP} is based on figures reported in [16, 128]. For our evaluation, foundry latency starts at 12 weeks for legacy nodes and increases starting from 20nm up to 20 weeks for 5nm; packaging latency is set to 6 weeks for all process nodes.

To find die area $A_{die}(d, p)$, we use a design's transistor count and available/estimated transistor densities at each process node from [24, 54]. The number of wafers $N_W(p)$ is found from the final number of chips multiplied by the die area divided by the wafer area. Our model also accounts for partial edge dies. In our evaluation, all results are calculated using 300mm diameter equivalent

[1] The parameters used in this work are selected to be representational of current conditions, but should **not** be interpreted to be actual figures for any chip design firms, EDA tools, process nodes, foundries, backend services, OSAT firms, *etc.*

Table 2: Estimated Wafer Production Rates Across Process Nodes [118]

Process Node	250nm	180nm	130nm	90nm	65nm	40nm	28nm	20nm	14nm	10nm	7nm	5nm
Wafer Production Rate (kWafer/Month)	41	241	120	79	189	284	350	0	281	0	252	97

wafers (some legacy process nodes are still fabricated on 200mm wafers [66]).

Foundry demand and queuing times for process nodes are difficult to estimate for current supply chain conditions, as almost all sources have reported lead times in aggregate [58, 65, 68, 70, 71]. In our evaluation, we assume the queuing time $T_{fab,queue}$ to be zero unless specified, meaning the time-to-market numbers reflect the most optimistic estimates for the architecture, process node, and final chip quantity. All time values reported in our evaluation are in calendar weeks.

Chip creation cost modeling includes both tapeout engineering costs and manufacturing costs and is adopted from Moonwalk [56]. We augmented Moonwalk's modeling to include new process nodes, manufacturing packaging costs, and updated mask costs [50].

We incorporate variance-based sensitivity sampling and analysis to study how our model's results behave from input variance [107]. We analyze six inputs that are difficult to estimate since they are closely guarded by foundries and design firms: defect density, wafer production rate, foundry latency, OSAT latency, total transistor count, and unique transistor count. We vary the six inputs with a ±10% error range from our estimates. Our time-to-market and *CAS* report the average of 1024 samples. Additionally, we report the 95% confidence interval (CI) for output variance given a ±10% and ±25% input variance, which is shown as pink and green error bars in Figures 7 and 11 and shaded regions in Figures 9 and 12.

Due to the desire to maintain their competitive advantage, companies involved in the chip fabrication process maintain the majority of information regarding their operation private or under non-disclosure agreements. This makes the verification of the absolute values of our model impossible. The chip creation model provides a useful first-order approximation for quickly evaluating chip designs in a time-to-market and manufacturing supply chain context. Even if the true values of the parameters were available, market parameters are constantly changing - therefore the discussion of our model will focus on the relative results that can be used for comparing designs.

6 REAL WORLD CASE STUDIES

In this section, we evaluate our model through five case studies of "real world" designs.

6.1 Case Study: Cache Sizing

Architects need to optimize designs not only for performance goals but also for market conditions. For chips where performance is paramount or when a design will be sold for a long time, architects may only need to focus on performance metrics such as IPC. However, during a severe manufacturing crisis or in a highly competitive market where being the first to release a chip is crucial, architects must consider factors such as cost and time-to-market.

For example, computer architects can increase performance by increasing hardware component capacity (*e.g.* caches, branch history table, core count). However, increasing hardware components

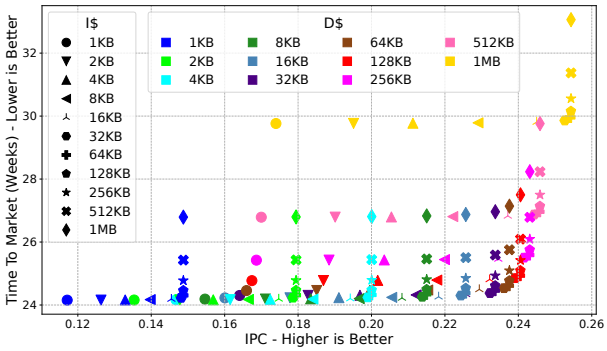

Figure 4: IPC and Time-to-Market for (I$, D$) Capacity When Manufacturing 100M 16 Core Ariane Chips Manufactured at 14nm. Each marker and color type represented a fixed I$ and D$ capacity respectively.

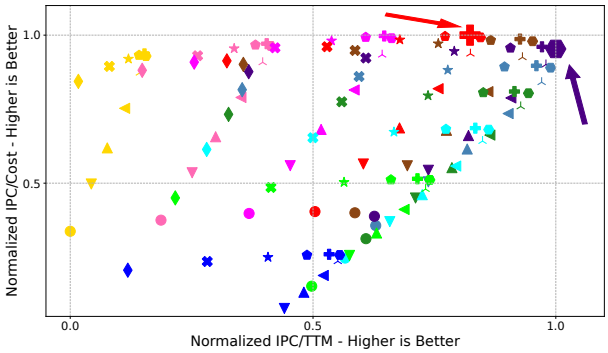

Figure 5: Normalized IPC/TTM and IPC/Cost for (I$, D$) Capacity. The purple arrow denotes the highest IPC/TTM and red arrow denotes the highest IPC/cost

for performance can increase time-to-market and cost: with caches, increasing IPC would favor a larger cache to reduce cache misses but increasing chip area will require more wafers and decrease chip yield, leading to longer fabrication times and manufacturing costs.

Architects often use performance per cost to realize performance at the best value. By utilizing this work's time-to-market model, architects can now calculate performance per time-to-market (*e.g.* IPC per Week) to find designs that maximize runtime performance and also minimize design, tapeout, and manufacturing time.

As an example to show how our time-to-market model can help architects and influence computer architecture, we present a case study that evaluates how cache sizes influence performance, cost, and time-to-market. We show that optimizing for performance per cost is **not optimal** when designing for a mass-produced chip that must have high performance relative and short time-to-market. We use Ariane's [129] architecture (originally with 16 KB instruction cache and 32 KB data cache) and SPEC2000 cache performance metrics [18] which sweep instruction cache and data cache from 1KB to 1MB to create an IPC model.

Figure 4 shows a scatter plot of IPC and time-to-market for each (instruction cache, data cache) pair for manufacturing 100M 16

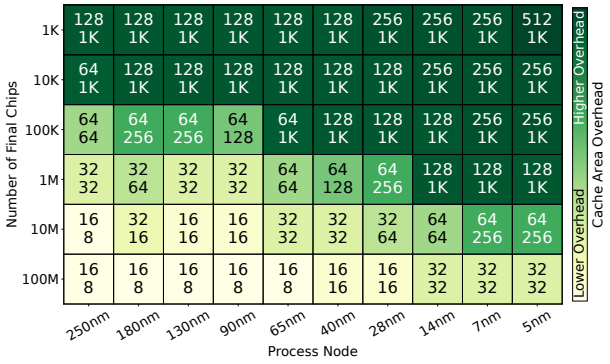

Figure 6: IPC/TTM Optimized Instruction and Data Cache $\binom{I\$}{D\$}$ Configurations (KB) for 16 Core Ariane for Select Process Nodes and Chips Manufactured. The color bar represents the cache area overhead relative to total die area.

core Ariane chips at 14nm. When cache sizes are small, doubling the cache capacity can significantly increase IPC with little time-to-market trade-offs. However, as total cache sizes increase over 512KB, we see diminishing performance improvements and the increased die area pushes up time-to-market.

Figure 5 shows a scatter plot of IPC/TTM on the x-axis and IPC/cost on the y-axis for the previous configuration. Although both IPC/TTM and IPC/Cost exhibit diminishing returns past a certain cache size, the optimal configuration is different between the two metrics. IPC/TTM is maximized with 32 KB instruction cache and data cache (denoted by purple hexagon) while IPC/cost peaks with a 64 KB instruction cache and 128 KB data cache (denoted by red plus). Optimizing for IPC/TTM provides better overall performance, time-to-market, and cost compared to IPC/cost. While the IPC/TTM optimal (32 KB, 32 KB) design's IPC/cost is 4% less than the max IPC/cost, the IPC/cost optimal (64 KB, 128 KB) design's IPC/TTM is 18% less than the max IPC/TTM. This analysis is crucial if there is a race to release a product and shows the importance of time-to-market analysis in a competitive market.

Figure 6 shows instruction and data cache configurations that maximize IPC/TTM for select process nodes and chips produced. As process nodes shrinks, the cache's area costs become less significant as more dies can be manufactured from a single wafer, therefore IPC/TTM favors increasing cache capacity to increase IPC. Furthermore, as chip quantity increases, the wafer production rate becomes the bottleneck over foundry latency; a smaller chip area becomes more important in order to reduce the number of ordered wafers. Generally, larger data caches are preferred compared to instruction caches. However, for large-scale production at legacy nodes, larger instruction caches than data caches are preferred when optimizing for performance and time-to-market. These insights highlight the importance of time-to-market analysis in a competitive market.

6.2 Case Study: Apple A11

In their 2022 Q2 quarterly earnings results, TSMC reported generating 0% of its revenue from their 20nm and 10nm process nodes [118], which likely indicates little to no current production at these processes. Considering the high foundry demand across all process nodes and slow ramp-up times, there would be an extended delay

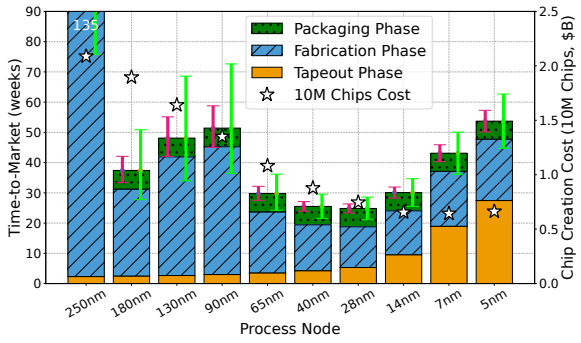

Figure 7: Time-to-Market (Left Axis) and Cost (Right Axis) for 10 million A11 chips - Lower is Better. Pink and green error bars represent the output variance's 95% CI under ±10% and ±25% input variance respectively.

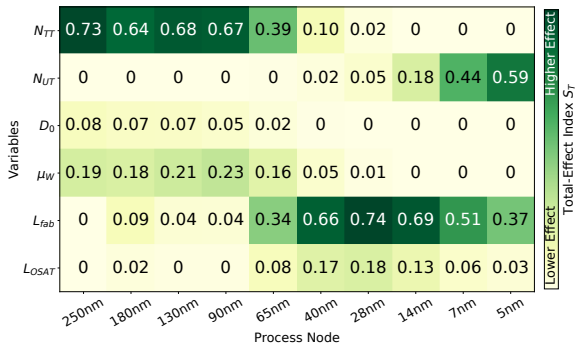

Figure 8: Sensitivity Analysis for Time-to-Market for 10 million A11 chips by Process Node - Higher S_T indicates more influence on output variance

before they could start production again at 20nm and 10nm. The Apple A11 processor was one of the first chips manufactured on TSMC's 10nm process [35]. If the architecture needed to be produced again now, the chip creation process would have to restart from the tapeout phase at a new process node. We use our modeling framework to estimate time-to-market, cost, and agility to re-create A11 chips now using expected market conditions.

The known components of the A11 chip architecture are: two big and four little CPU cores, three GPU cores, and a neural processing unit; all blocks are custom designed in house by Apple [35]. The chip has a die area of 88mm^2 at 10nm and uses 4.3 billion transistors.

We consider the block area estimates of the big CPU, little CPU, GPU cores, and the NPU reported by [35] multiplied by our transistor density model for 10nm to estimate unique transistor count N_{UT}. We assume the rest of the die area is populated by memory and other soft IP that have been pre-verified at gate-level and are available for every process node from third-party vendors. We estimate N_{UT} to be ~ 514 million transistors. The total engineering-weeks to complete the tapeout phase for each process node is converted to calendar weeks by assuming a team of 100 tapeout engineers and each individual block can be done in parallel and then synchronized for the top-level tapeout.

Figure 7 shows the time-to-market and cost for producing 10 million final chips across different process nodes. For our example's

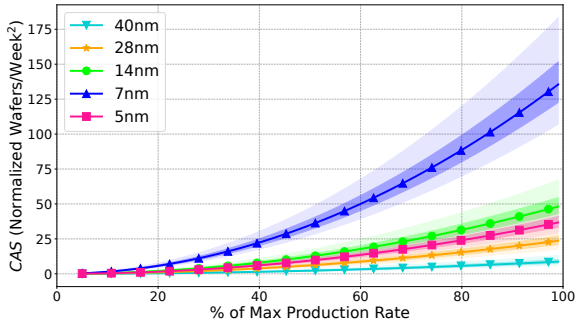

Figure 9: *CAS* for 10 million A11 chips - Higher is Better. The light and dark shaded regions represent the output variance's 95% CI under ±10% and ±25% input variance respectively.

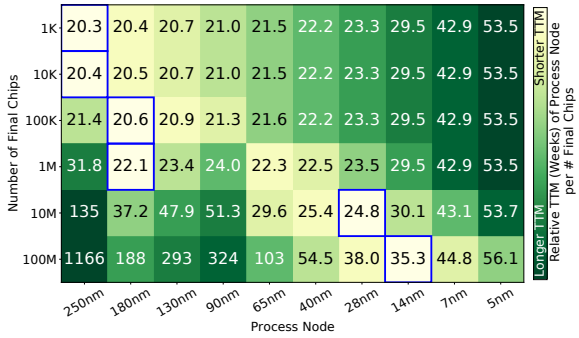

Figure 10: Time-to-market Matrix for A11 chips - Lower is Better. The fastest process for each number of final chips is outlined in blue.

A11 chip architecture, the 28nm process has the quickest time-to-market. The 250, 130, and 90nm processes require long fabrication phases due to low transistor density and low wafer production - a 4.3 billion transistor chip at the 250nm process node would only fit 43 dies per 300mm wafer with an expected 48% die yield. This in turn requires more wafers to be purchased which dominates chip creation costs. Similarly, more advanced nodes have higher tapeout costs but require fewer wafers which leads to lower overall costs. The 14nm process has a similar fabrication/packaging time compared to 28nm but has a longer time-to-market due to longer expected tapeout time despite requiring 3.16x fewer wafers compared to 28nm. The 5nm process node has a longer fabrication phase compared to 7nm and 14nm due to its lower wafer production rate despite requiring 1.84x and 6.44x fewer wafers compared to 7nm and 14nm respectively.

Figure 8 shows the A11 time-to-market sensitivity analysis for the six varied input parameters. For the 250nm through 90nm processes, the majority of output variance can be attributed to the total transistor count. This is because chip area is more sensitive at lower transistor densities combined with lower wafer production rates. Between 65nm and 7nm, foundry and OSAT latencies begin to dominate time-to-market variance as they bottleneck the higher good chips per week at these process nodes. The 5nm process' time-to-market is most affected by unique transistor count which reflects the exponentially increasing tapeout costs and efforts at the most advanced process nodes.

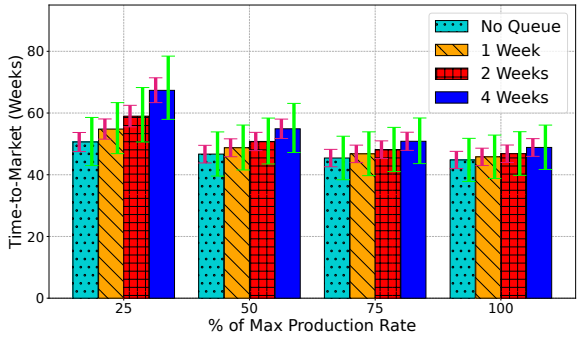

Figure 11: Time-to-Market for 10 million A11 chips at 7nm by $T_{fab,queue}$ - Lower is Better. Pink and green error bars represent the output variance's 95% CI under ±10% and ±25% input variance respectively.

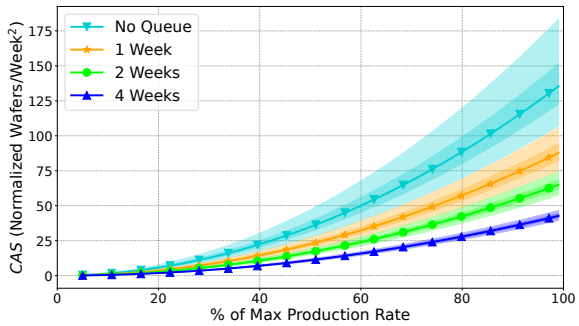

Figure 12: *CAS* for 10 million A11 chips at 7nm using $T_{fab,queue}$ - Higher is Better. The light and dark shaded regions represent the output variance's 95% CI under ±10% and ±25% input variance respectively.

To assess the agility of this architecture, we computed the *CAS* for our modeled A11 at the five most advanced process nodes, shown in Figure 9. The 7nm process has the highest agility due to its high wafer production rate and transistor density despite the higher time-to-market. Similarly, the lower densities and production rates at 40nm and 28nm lead to lower *CAS* compared to more advanced nodes. The higher transistor densities at 5nm mean an equal change in wafer production rate will result in a greater change in overall die production rate compared to larger process nodes. Combined with a lower overall wafer production rate, the A11 chip architecture is more agile at 14nm compared to 5nm for producing 100 million chips with current market conditions.

The fastest time-to-market depends on the process node as well as the number of final chips produced. The time-to-market of the A11 architecture is evaluated for select quantities of final chips and results are shown in Figure 10. As the number of chips increases, the time-to-market generally shifts towards more advanced process nodes, as higher transistor densities and production rates begin to outweigh tapeout time. Under current market conditions, the 180nm process is faster to market than 130 and 90nm processes even up to 100M chips due to a higher wafer production rate.

Table 3: Accelerator Speed-Up ($\frac{cycles}{cycles}$), $T_{tapeout}$ (weeks), and $C_{tapeout}$

Hardware Block	Speed-Up	N_{TT}	Area Relative to Ariane	$T_{tapeout}$ ($5nm$)	$C_{tapeout}$ ($5nm$)
Sorting Stream	16.71x	45.62M	18.18x	3.5	$6.8M
Sorting Iterative	3.07x	18.90M	7.53x	1.6	$4.6M
DFT Stream	56.36x	37.31M	14.87x	2.9	$6.1M
DFT Iterative	20.81x	18.18M	7.24x	1.5	$4.6M

6.3 Case Study: Foundry Demand and $T_{fab,queue}$

One of the main causes of the current chip shortage is the sudden increase in foundry demand which has increased queuing time $T_{fab,queue}$ but not affected production rates. In order to explore the effects of increased queuing time in the face of production interruptions, we evaluate time-to-market and *CAS* for the same A11 chip architecture produced at the 7nm process node for 10 million final chips and varying queuing time from 0 to 4 weeks. The results for time-to-market and *CAS* are shown in Figures 11 and 12 respectively.

As the wafer production rate decreases, it takes longer to fabricate all the wafers needed for the design as well as for the wafers ahead in the queue. If the foundry quotes an initial lead time based on the number of wafers it needs to complete, a severe production side disruption can steeply increase time-to-market and reduce the agility of the architecture - just 1 week of queue time decreased the maximum *CAS* by 37%. This furthers the importance of understanding the full context of current and future market conditions when designing chips to avoid time-to-market delays.

6.4 Case Study: Cost of Specialization

Assume that a design team already has a general-purpose core that is ready for tapeout and wants to improve their chip by incorporating an accelerator block. While accelerators provide performance and energy improvements, this change will incur additional tapeout time which delays the chip's time-to-market and increases tapeout costs $C_{tapeout}$.

As an example to evaluate these tradeoffs, we benchmark SPIRAL generated fixed point sorting [130] and floating point FFT accelerators [79] against a general purpose open-source core (Ariane [129]) running 2048 sized blocks for their respective tasks. Unique transistor counts come from synthesizing RTL designs with commercial EDA tools assuming that non-memory transistors are unique. Results for 5nm accelerators are shown in Table 3 to represent worst-case scenarios.

Even though accelerators show impressive speed-up over Ariane, they can increase $T_{tapeout}$ by almost a month and add millions USD to tapeout costs at 5nm. Results also show that streaming accelerators improve speed-up but tradeoff with increased tapeout time and costs. For routines that are parallelizable or partitionable, a uniform manycore design that can be quickly taped out may be more prudent than taping out an accelerator. During chip crises and economic downturns, understanding the tapeout tradeoffs of integrating specialized hardware helps architects optimize decisions to meet their performance, time-to-market, and cost requirements.

Table 4: Die Transistor Count, $T_{tapeout}$ (weeks), and Area for Zen 2-Like Architecture [86, 105]

Die	N_{TT}	N_{UT}	A_{die} (mm^2) ($14nm/7nm$)	$T_{tapeout}$ ($14nm/7nm$)
Compute	3.8B*	475M*	206 / 74*	3.6 / 10.4
I/O	2.1B*	523M	125* / 38	4.0 / 11.5

6.5 Case Study: Chiplets and Mixed-Process Nodes

With the rise of chiplet architectures, multiple distinct dies may be packaged in a single final chip [85]. Although desirable for their smaller dies and accordingly higher yields, they need to wait for the latest fabricated die before they can be packaged. Modern chiplet architectures may also combine logic dies and/or memory dies from different process nodes [15]. For example, AMD's Zen 2 microarchitecture is composed of compute dies fabricated at a 7nm process and central I/O dies fabricated at a 12nm process and integrated together in the same packaging. [86].

Some chiplet designs require silicon interposers, often created at legacy nodes [90]. If there is high demand and/or low production capacity at the logic's or interposer's nodes, the packaging process is delayed until both are produced. While some interposers may be as simple as routing metal layers (passive), others may contain complex features such as active transistor logic which means interposers also have to complete the chip creation process before being packaged with the logic dies [51, 90, 111].

To evaluate the time-to-market and *CAS* implications of chiplets and mixed-process node designs, we use a Zen 2-inspired chip architecture with two compute dies and one central I/O die. Transistor counts and die areas come from [86, 105], N_{UT} of the compute die is based on the transistor count for a single compute core, and N_{UT} of the I/O die is estimated to be 25% of I/O die transistor count from enthusiasts die photo annotations [115]. A summary of transistor counts, die area, and tapeout times can be found in Table 4; asterisks (*) indicate numbers are directly from [86, 105]. For interposer designs, the interposer is fabricated at the 65nm process [90] and is 120% the area of the chiplets packaged. We assume a passive interposer with an optimistic 99.99% yield.

We then calculate time-to-market, cost, and *CAS* based on the original Zen 2 design (one 12nm I/O die and two 7nm compute dies with no interposer) and compare with hypothetical designs of Zen 2 with interposer, all three chiplets at 7nm with and without interposer, all three chiplets at 12nm with and without interposer, a 7nm monolithic equivalent, and a 12nm monolithic equivalent. Results are shown in Figure 13.

The original Zen 2 design has a faster time-to-market (Fig. 13a) compared to an all 7nm design, as the compute and I/O dies can be produced in parallel. Additionally, the mixed-process design spends fewer engineering-hours in the tapeout phase and once the 12nm I/O design finishes its tapeout, it can move forward to the fabrication phase independent of the 7nm compute die. However, mixed-process designs cost more (Fig. 13b) as two processes contribute to tapeout and manufacturing costs. Our results also demonstrate that chiplet designs without interposers have faster time-to-market, are more cost-efficient, and achieve higher agility compared to equivalent monolithic designs.

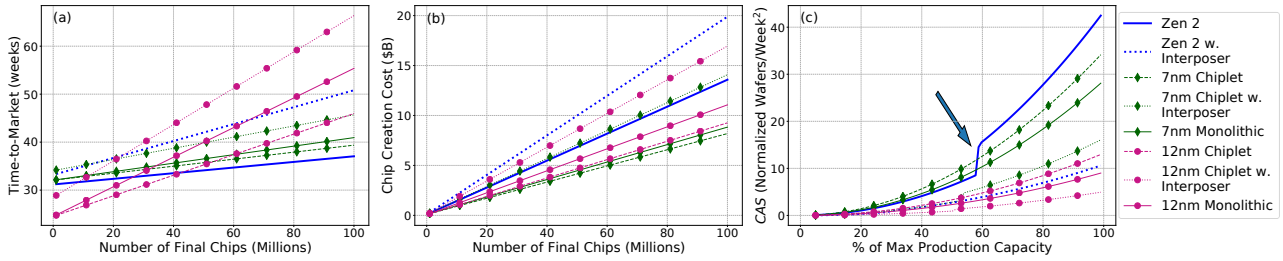

Figure 13: (a) Time-to-Market, (b) Chip Creation Cost, and (c) *CAS* for Mixed-Process Chiplet Study

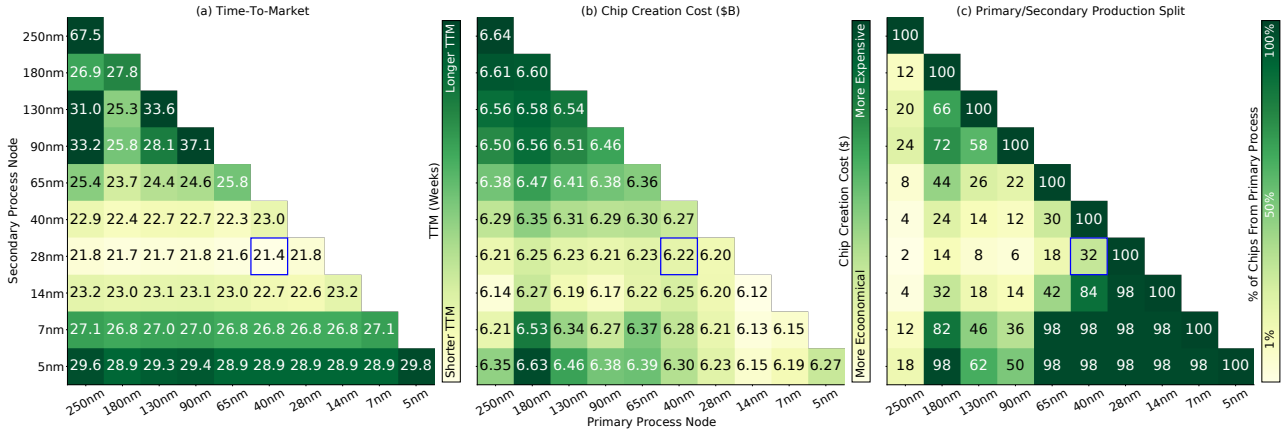

Figure 14: (a) Time-to-market, (b) Chip Creation Cost, and (c) Production Split for Two Process Chip Design Study - overall fastest time-to-market's combination is highlighted in blue

Chiplet designs that require interposers have the worst time-to-market, cost, and *CAS* metrics, as interposers require large die area and are produced at a lower capacity legacy process node. In a supply-constraint market, the marginal increases in chiplet yields may be offset by the lead times for interposer production. For chiplet-interposer architectures to succeed in the face of current supply chain disruptions, interposers must be accommodating to being built on different process nodes. Fabricating the interposer at the higher-wafer-production-rate 40nm process decreases time-to-market for 100 million final chips from 51 weeks to 45 weeks and increases max *CAS* by 126% with only a $77M increase in chip creation costs.

The final number of chips is an important factor when comparing process nodes. The time-to-market is faster when producing fewer final chips at the 12nm process due to lower tapeout time and lower fab latency compared to 7nm. As the number of final chips increases, the 7nm process' higher good chips per week eventually outpaces the 12nm's.

As seen in Figure 13c, the original design has the highest *CAS* at full production capacity (right of the plot) but reaches a steep decline at 60% of max production rate (indicated by the arrow) and drops to below the 7nm chiplet monolithic design's blue *CAS* curves at low wafer production capacity (left of the plot). This behavior is explained by the synchronization step at the packaging phase. When both 7nm and 12nm process nodes are at full capacity, the 12nm I/O dies complete the fabrication phase first and must wait on the 7nm dies before beginning packaging. Therefore when there is only a small decrease in production rate on the 12nm process, only the 7nm process contributes to the agility score which is higher compared to the other 7nm designs since fewer wafers are needed. Once below the 60% max production rate for the 12nm process, it becomes the production stage bottleneck and lowers the *CAS*, demonstrating how mixed process node chip architectures incur additional vulnerability to production rate changes.

7 MULTI-PROCESS CHIP MANUFACTURING

As shown in Section 6.5, splitting up chip manufacturing over multiple process nodes can decrease time-to-market and increase *CAS* but increases chip creation costs. In this section, we propose a chip design methodology that tapes out the same architecture on two process nodes in parallel to utilize the combined manufacturing capacity. We use time-to-market, cost, and *CAS* modeling to find optimal production splits.

Previous industry and academic chips have ported the same architecture across multiple process nodes: Sony/IBM's Cell (90, 65, 45nm) [103, 110]; Nvidia's Tegra X1 (20, 16nm) [43]; Ariane (22, 4nm) [19, 129]; Rocket chip (45, 28, 22nm) [10], etc. Industry chips that change process nodes can rerun the tapeout after the initial release, but may still market the product under the same SKUs (seen with PlayStation 3 and Nintendo Switch gaming consoles). Designers may not necessarily want to improve performance once ported to a new process - some applications require bug-for-bug compatibility and modifications may introduce new defects.

To explore our proposed methodology, we modeled the Raven/PicoRV32 architecture [28], which has been previously taped out on 180nm. The performance and chip area are akin to a low-end ARM Cortex-M IP commonly used in automotive and cross-market microcontrollers [109]. The minimum die area is set to 1 mm^2.

We sweep each process node pair (called primary and secondary) and find the chip production split that has the highest *CAS* for a multicore Raven-inspired design for 1 billion final chips. Results are shown in Figure 14.

When optimizing for highest *CAS*, a combination of the 28nm and 40nm processes has the fastest time-to-market (and also the highest agility) due to their highest wafer production capacities. The longer foundry latency penalizes advanced process nodes and *CAS* optimization pushes the majority of chips to be produced on the more legacy node. By leveraging two processes to produce equivalent chips in parallel, time-to-market can be reduced compared to a single process design. For legacy nodes that have lower wafer production rates (eg. 250, 130, 90nm), adding parallel manufacturing on the next smaller process can save 40, 6, and 13 weeks off time-to-market respectively.

Without mass production, it is difficult for architects to justify the additional tapeout cost of multi-process chip design. In this example, the relatively low transistor count means tapeout time and costs are outweighed by fabrication and packaging, making designing chips for multiple processes economically feasible. Using two legacy node processes may also reduce overall costs. The transistor density improvements of the next process node mean fewer total wafers are needed which offsets the additional tapeout and mask costs.

For mass-produced, legacy node compatible chips, using two process nodes can decrease time-to-market and be more cost-efficient. By maximizing *CAS*, the optimal production split reflects high capacity and agile processes that help architects design supply chain resilient chips.

8 RELATED WORK

Previous work in computer architecture that studied the chip creation process in the context of chip architecture and transistor scaling has mainly focused on optimizing for cost [56, 78], performance [36, 83], and energy [29, 42, 122]. In contrast, and complementary, to these works, our work focuses on how a chip's architecture affects its time-to-market and agility to weather supply chain disruptions.

A significant amount of research has explored ways to speedup portions or the entire chip creation process, especially with a focus on the design and tapeout phases. Estimators and open-source tools help architects assess design tradeoffs [12, 21, 52, 101] and quickly generate RTL for tapeout [10, 13, 72]. Other architecture and EDA works focus on automating the layout process, reducing human engineering effort, and optimizing the place-and-route process [5, 8, 61, 92]. Design for testing and design for manufacturing principles inspire chip architectures that require less testing and packaging effort. The impact of these speedups are independent of this work and can be incorporated into the parameters of the chip creation model and be applied in parallel to help computer architects create and receive their chips faster.

Supply chain management research has previously investigated the resilience and agility of supply chains and we adapt the agility terminology in a computer architecture context [4, 120]. Furthermore, comprehensive models of the semiconductor fabrication supply chain have been developed, but they do not involve the design nor tapeout phases and lack the ability to assess how individual chip architectures affect the duration of the fabrication phase [55, 75, 84]. This work combines the fields of supply chains, semiconductor manufacturing, and computer architecture to derive architectural insights to design supply chain aware chips.

9 CONCLUSION

In this work, we investigated how computer architecture design choices can enable chips to be more resilient to semiconductor manufacturing supply chain disruptions. We are the first work to make time-to-market and supply chain conditions a first-class design constraint in computer architecture research and chip design.

We characterized and modeled the steps of the chip creation process and provided a model for architects to quickly assess the time-to-market of their chips depending on its architecture and current market conditions. In addition, we propose the Chip Agility Score (*CAS*) to evaluate a chip architecture's agility in the face of production side supply chain interruptions. *CAS* enables architects to evaluate their architectures against expected market conditions to design chips that are more resilient to chip shortages.

Our modeling framework and data sets are open-sourced to advance supply chain aware computer architecture research. https://github.com/PrincetonUniversity/ttm-cas

ACKNOWLEDGMENTS

We would like to thank Grigory Chirkov, Marcelo Orenes-Vera, Rohan Prabhakar, as well as the entire Princeton Parallel Group, and our anonymous reviewers for their feedback, suggestions, and encouragement. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2039656, the National Science Foundation under Grant No. CNS-1823222, Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) under agreement No. FA8650-18-2-7862. 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. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) or the U.S. Government.

REFERENCES

[1] 2015. International Technology Roadmap for Semiconductors 2.0 2015 Edition Test and Test Equipment. (Jun 2015).
[2] Achronix. 2022. VectorPath Accelerator Card. https://www.achronix.com/product/vectorpath-accelerator-card
[3] Peter J. Adams, Brannon Batson, Alistair Bell, Jhanvi Bhatt, J. Adam Butts, Timothy Correia, Bruce Edwards, Peter Feldmann, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Maria Gorlatova, Brian Greskamp, J.P. Grossman, Jeremy Hunt, Bryan L. Jackson, Mollie M. Kirk, Jeffrey S. Kuskin, Roy J. Mader, Richard McGowen, Adam McLaughlin, Mark A. Moraes, Mohamed Nasr, Lawrence J. Nociolo, Lief O'Donnell, Andrew Parker, Jon L. Peticolas, Terry Quan, T. Carl Schwink, Keun Sup Shim, Naseer Siddique, Jochen Spengler, Michael Theobald, Brian Towles, William Vick, Stanley C. Wang, Michael Wazlowski, Madeleine J. Weingarten, John M. Williams, and David E.Shaw. 2021. The ANTON 3 ASIC: a Fire-Breathing Monster for Molecular Dynamics Simulations. In *2021 IEEE Hot Chips 33 Symposium (HCS)*. 1–22.

https://doi.org/10.1109/HCS52781.2021.9567084

[4] Ashish Agarwal, Ravi Shankar, and M.K. Tiwari. 2007. Modeling agility of supply chain. *Industrial Marketing Management* 36, 4 (2007), 443–457. https://doi.org/10.1016/j.indmarman.2005.12.004

[5] T Ajayi, D Blaauw, TB Chan, CK Cheng, VA Chhabria, DK Choo, M Coltella, S Dobre, R Dreslinski, M Fogaça, S. Hashemi, A. Hosny, A. B. Kahng, Minsoo Kim, J. Li, Z. Liang, U. Mallappa, P. Penzes, G. Pradipta, S. Reda, A. Rovinski, K. Samadi, S. S. Sapatnekar, L Saul, C. Sechen, V. Srinivas, W. Swartz, D. Sylvester, D. Urquhart, L. Wang, M. Woo, and B. Xu. 2019. OpenROAD: Toward a Self-Driving, Open-Source Digital Layout Implementation Tool Chain. *Proc. GOMACTECH* (2019), 1105–1110.

[6] Alexandra Alper. 2022. Russia's attack on Ukraine halts half of world's neon output for chips. *Reuters* (Mar 2022). https://www.reuters.com/technology/exclusive-ukraine-halts-half-worlds-neon-output-chips-clouding-outlook-2022-03-11/

[7] Alexandra Alper and Karen Freifeld. 2022. White House tells chip industry to be ready for potential Russia export curbs. *Reuters* (Jan 2022). https://www.reuters.com/business/white-house-tells-chip-industry-be-ready-potential-russia-export-curbs-2022-01-19/

[8] Alon Amid, David Biancolin, Abraham Gonzalez, Daniel Grubb, Sagar Karandikar, Harrison Liew, Albert Magyar, Howard Mao, Albert Ou, Nathan Pemberton, Paul Rigge, Colin Schmidt, John Wright, Jerry Zhao, Yakun Sophia Shao, Krste Asanović, and Borivoje Nikolić. 2020. Chipyard: Integrated Design, Simulation, and Implementation Framework for Custom SoCs. *IEEE Micro* 40, 4 (2020), 10–21. https://doi.org/10.1109/MM.2020.2996616

[9] Apple. 2020. Apple announces Mac transition to Apple silicon. https://www.apple.com/newsroom/2020/06/apple-announces-mac-transition-to-apple-silicon/

[10] Krste Asanović, Rimas Avizienis, Jonathan Bachrach, Scott Beamer, David Biancolin, Christopher Celio, Henry Cook, Daniel Dabbelt, John Hauser, Adam Izraelevitz, Sagar Karandikar, Ben Keller, Donggyu Kim, John Koenig, Yunsup Lee, Eric Love, Martin Maas, Albert Magyar, Howard Mao, Miquel Moreto, Albert Ou, David A. Patterson, Brian Richards, Colin Schmidt, Stephen Twigg, Huy Vo, and Andrew Waterman. 2016. *The Rocket Chip Generator.* Technical Report UCB/EECS-2016-17. EECS Department, University of California, Berkeley. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-17.html

[11] Semiconductor Industry Association. 2022. Global Semiconductor Sales, Units Shipped Reach All-Time Highs in 2021 as Industry Ramps Up Production Amid Shortage. https://www.semiconductors.org/global-semiconductor-sales-units-shipped-reach-all-time-highs-in-2021-as-industry-ramps-up-production-amid-shortage/

[12] Jonathan Balkind, Katie Lim, Michael Schaffner, Fei Gao, Grigory Chirkov, Ang Li, Alexey Lavrov, Tri M. Nguyen, Yaosheng Fu, Florian Zaruba, Kunal Gulati, Luca Benini, and David Wentzlaff. 2020. BYOC: A "Bring Your Own Core" Framework for Heterogeneous-ISA Research. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems* (Lausanne, Switzerland) *(ASPLOS '20)*. Association for Computing Machinery, New York, NY, USA, 699–714. https://doi.org/10.1145/3373376.3378479

[13] Jonathan Balkind, Michael McKeown, Yaosheng Fu, Tri Nguyen, Yanqi Zhou, Alexey Lavrov, Mohammad Shahrad, Adi Fuchs, Samuel Payne, Xiaohua Liang, Matthew Matl, and David Wentzlaff. 2016. OpenPiton: An Open Source Manycore Research Framework. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems* (Atlanta, Georgia, USA) *(ASPLOS '16)*. Association for Computing Machinery, New York, NY, USA, 217–232. https://doi.org/10.1145/2872362.2872414

[14] Pete Bannon, Ganesh Venkataramanan, Debjit Das Sarma, and Emil Talpes. 2019. Computer and Redundancy Solution for the Full Self-Driving Computer. In *2019 IEEE Hot Chips 31 Symposium (HCS)*. 1–22. https://doi.org/10.1109/HOTCHIPS.2019.8875645

[15] Bryan Black. 2013. Die Stacking is Happening Keynote. In *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture* (Davis, California) *(MICRO-46)*. Association for Computing Machinery, New York, NY, USA. https://www.microarch.org/micro46/files/keynote1.pdf

[16] Andrew Blum. 2021. From Cars to Toasters, America's Semiconductor Shortage Is Wreaking Havoc on Our Lives. Can We Fix It? *Time* (Jun 2021). https://time.com/6075425/semiconductor-chip-shortage

[17] Ondrej Burkacky, Julia Dragon, and Nikolaus Lehmann. 2022. The semiconductor decade: A trillion-dollar industry. https://www.mckinsey.com/industries/semiconductors/our-insights/the-semiconductor-decade-a-trillion-dollar-industry

[18] Jason F. Cantin and Mark D. Hill. 2001. Cache Performance for Selected SPEC CPU2000 Benchmarks. *SIGARCH Comput. Archit. News* 29, 4 (sep 2001), 13–18. https://doi.org/10.1145/563519.563522

[19] Gregory K. Chen, Phil C. Knag, Carlos Tokunaga, and Ram K. Krishnamurthy. 2022. An 8-core RISC-V Processor with Compute near Last Level Cache in Intel 4 CMOS. In *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. 68–69. https://doi.org/10.1109/VLSITechnologyandCir46769.2022.9830518

[20] Ting-Fang Cheng and Lauly Li. 2022. TSMC struggles to keep new hires, warns of power supply risks. *Nikkei Asia* (June 2022). https://asia.nikkei.com/Business/Tech/Semiconductors/TSMC-struggles-to-keep-new-hires-warns-of-power-supply-risks

[21] Grigory Chirkov and David Wentzlaff. 2023. SMAPPIC: Scalable Multi-FPGA Architecture Prototype Platform in the Cloud. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2* (Vancouver, BC, Canada) *(ASPLOS 2023)*. Association for Computing Machinery, New York, NY, USA, 733–746. https://doi.org/10.1145/3575693.3575753

[22] Don Clark. 2021. The Tech Cold War's 'Most Complicated Machine' That's Out of China's Reach. *The New York Times* (July 2021). https://www.nytimes.com/2021/07/04/technology/tech-cold-war-chips.html

[23] Gabrielle Coppola, Tara Patel, and Debby Wu. 2021. Chip Shortage Forces Carmakers to Leave Out Some High-End Features. *Bloomberg* (May 2021). https://www.bloomberg.com/news/articles/2021-05-06/chip-shortage-forces-carmakers-to-strip-out-high-tech-features

[24] Rachel Courtland. 2017. Intel Now Packs 100 Million Transistors in Each Square Millimeter. *IEEE Spectrum* (Mar 2017). https://spectrum.ieee.org/intel-now-packs-100-million-transistors-in-each-square-millimeter

[25] Mike Cronin. 2021. Samsung announces chip manufacturing has resumed in Austin. *Austin Business Journal* (March 2021). https://www.bizjournals.com/austin/news/2021/03/30/samsung-austin-semiconductor-is-back-online.html

[26] J.A. Cunningham. 1990. The use and evaluation of yield models in integrated circuit manufacturing. *IEEE Transactions on Semiconductor Manufacturing* 3, 2 (1990), 60–71. https://doi.org/10.1109/66.53188

[27] Ian Cutress. 2020. 'Better Yield on 5nm than 7nm': TSMC Update on Defect Rates of N5. *AnandTech* (Aug 2020). https://www.anandtech.com/show/16028/better-yield-on-5nm-than-7nm-tsmc-update-on-defect-rates-for-n5

[28] efabless engineering. [n. d.]. Raven: An ASIC implementation of the PicoSoC PicoRV32. https://github.com/efabless/raven-picorv32

[29] Hadi Esmaeilzadeh, Emily Blem, Renée St. Amant, Karthikeyan Sankaralingam, and Doug Burger. 2011. Dark silicon and the end of multicore scaling. In *2011 38th Annual International Symposium on Computer Architecture (ISCA)*. 365–376.

[30] Jack Ewing. 2022. Why Tesla Soared as Other Automakers Struggled to Make Cars. *The New York Times* (Jan. 2022). https://www.nytimes.com/2022/01/08/business/teslas-computer-chips-supply-chain.html

[31] Asa Fitch. 2021. Texas Winter Storm Strikes Chip Makers, Compounding Supply Woes. *The Wall Street Journal* (Feb. 2021). https://www.wsj.com/articles/texas-winter-storm-strikes-chip-makers-compounding-supply-woes-11613588617

[32] Asa Fitch. 2022. Chip Makers Stockpiled Key Materials Ahead of Russian Invasion of Ukraine. *The Wall Street Journal* (Mar 2022). https://www.wsj.com/articles/chip-makers-stockpiled-key-materials-ahead-of-russian-invasion-of-ukraine-11647167582

[33] Samsung Foundry. 2022. About Samsung Foundry. https://www.samsungfoundry.com/foundry/homepage/anonymous/about.do?_mainLayOut=homepageLayout&menuIndex=01

[34] Samsung Foundry. 2022. Samsung Foundry Advanced Package Technology. https://www.samsungfoundry.com/foundry/homepage/anonymous/technologyAdvanced_new.do?_mainLayOut=homepageLayout&menuIndex=0203

[35] Andrei Frumusanu. 2018. The iPhone XS & XS Max Review: Unveiling the Silicon Secrets. *AnandTech* (Oct 2018). https://www.anandtech.com/show/13392/the-iphone-xs-xs-max-review-unveiling-the-silicon-secrets

[36] Adi Fuchs and David Wentzlaff. 2019. The Accelerator Wall: Limits of Chip Specialization. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 1–14. https://doi.org/10.1109/HPCA.2019.00023

[37] Thomas Fuller. 2011. Thailand Flooding Cripples Hard-Drive Suppliers. *The New York Times* (Nov 2011). https://www.nytimes.com/2011/11/07/business/global/07iht-floods07.html

[38] GlobalFoundries. 2022. GF Post Fab Services. https://gf.com/gf-post-fab-services

[39] GlobalFoundries. 2022. STMicroelectronics and GlobalFoundries to advance FD-SOI ecosystem with new 300mm manufacturing facility in France. https://gf.com/gf-press-release/stmicroelectronics-and-globalfoundries-to-advance-fd-soi-ecosystem-with-new-300mm-manufacturing-facility-in-france/

[40] Samuel Goodman, John VerWey, and Dan Kim. 2019. The South Korea-Japan Trade Dispute in Context: Semiconductor Manufacturing, Chemicals, and Concentrated Supply Chains. In *Office of Industries Staff Publications and Research Papers (Office of Industries Working Paper ID-062)*. United States International Trade Commission, Washington, DC, USA, 1–36. https://doi.org/10.2139/ssrn.3470271

[41] Monika Gupta. 2021. Google Tensor is a milestone for machine learning. https://blog.google/products/pixel/introducing-google-tensor/

[42] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. 2022. ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tool. In *Proceedings of the*

49th Annual International Symposium on Computer Architecture (New York, New York) *(ISCA '22)*. Association for Computing Machinery, New York, NY, USA, 784–799. https://doi.org/10.1145/3470496.3527408

[43] Hilbert Hagedoorn. 2020. NVIDIA Shield Android TV 2019. *The Guru of 3D* (March 2020). https://www.guru3d.com/articles_pages/nvidia_shield_android_tv_2019_review,2.html

[44] HiSilicon. 2022. HiSilicon Kirin Chipsets. https://www.hisilicon.com/en/products/Kirin

[45] Intel. 2021. Intel Accelerated 2021. https://www.intel.com/content/www/us/en/events/accelerated.html

[46] Intel. 2022. Intel Announces Next US Site with Landmark Investment in Ohio. https://www.intel.com/content/www/us/en/newsroom/news/intel-announces-next-us-site-landmark-investment-ohio.html

[47] Eun-Young Jeong and Dan Strumpf. 2021. Why the Chip Shortage Is So Hard to Overcome. *The Wall Street Journal* (April 2021). https://www.wsj.com/articles/why-the-chip-shortage-is-so-hard-to-overcome-11618844905

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

[49] Handel Jones. 2014. Strategies In Optimizing Market Positions For Semiconductor Vendors Based On IP Leverage. *International Business Strategies* (Jan 2014).

[50] Scotten Jones. 2020. LithoVision – Economics in the 3D Era. *IC Knowledge* (March 2020). https://semiwiki.com/semiconductor-services/ic-knowledge/283426-lithovision-economics-in-the-3d-era/

[51] Ajaykumar Kannan, Natalie Enright Jerger, and Gabriel H. Loh. 2015. Enabling interposer-based disintegration of multi-core processors. In *2015 48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. 546–558. https://doi.org/10.1145/2830772.2830808

[52] Sagar Karandikar, Howard Mao, Donggyu Kim, David Biancolin, Alon Amid, Dayeol Lee, Nathan Pemberton, Emmanuel Amaro, Colin Schmidt, Aditya Chopra, Qijing Huang, Kyle Kovacs, Borivoje Nikolic, Randy Katz, Jonathan Bachrach, and Krste Asanović. 2018. FireSim: FPGA-accelerated Cycle-exact Scale-out System Simulation in the Public Cloud. In *Proceedings of the 45th Annual International Symposium on Computer Architecture* (Los Angeles, California) *(ISCA '18)*. IEEE Press, Piscataway, NJ, USA, 29–42. https://doi.org/10.1109/ISCA.2018.00014

[53] Wendy Kaufman. 2011. Thai Floods Disrupt Computer Hard Drive Supply. *NPR Morning Edition* (Nov 2011). https://www.npr.org/2011/11/25/142767696/thai-floods-disrupt-computer-hard-drive-supply

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

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

[56] Moein Khazraee, Lu Zhang, Luis Vega, and Michael Bedford Taylor. 2017. Moonwalk: NRE Optimization in ASIC Clouds. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems* (Xi'an, China) *(ASPLOS '17)*. Association for Computing Machinery, New York, NY, USA, 511–526. https://doi.org/10.1145/3037697.3037749

[57] Sohee Kim and Sam Kim. 2021. Korea Unveils $450 Billion Push for Global Chipmaking Crown. *Bloomberg* (May 2021). https://www.bloomberg.com/news/articles/2021-05-13/korea-unveils-450-billion-push-to-seize-global-chipmaking-crown

[58] Ian King. 2021. Chip Lead Times Begin to Slow, Suggesting Shortages Have Peaked. *Bloomberg* (Oct. 2021). https://www.bloomberg.com/news/articles/2021-10-26/chip-lead-times-begin-to-slow-suggesting-shortages-have-peaked

[59] Yoko Kubota. 2019. China Sets Up New $29 Billion Semiconductor Fund. *The Wall Street Journal* (Oct. 2019). https://www.wsj.com/articles/china-sets-up-new-29-billion-semiconductor-fund-11572034480

[60] Simon Kuo. 2021. Progress in Importation of US Equipment Dispels Doubts on SMIC's Capacity Expansion for Mature Nodes for Now. https://www.trendforce.com/presscenter/news/20210305-10693.html

[61] D.E. Lackey, P.S. Zuchowski, and J. Koehl. 2003. Designing mega-ASICs in nanogate technologies. In *Proceedings 2003. Design Automation Conference (IEEE Cat. No.03CH37451)*. 770–775. https://doi.org/10.1145/775832.776029

[62] Mark Lapedus. 2017. Battling Fab Cycle Times. *Semiconductor Engineering* (Feb. 2017). https://semiengineering.com/battling-fab-cycle-times/

[63] Mark Lapedus. 2018. Big Trouble At 3nm. *Semiconductor Engineering* (Jun 2018). https://semiengineering.com/big-trouble-at-3nm/

[64] Mark Lapedus. 2021. Challenges With Chiplets And Packaging. *Semiconductor Engineering* (Sep 2021). https://semiengineering.com/challenges-with-chiplets-and-packaging/

[65] Mark Lapedus. 2021. End In Sight For Chip Shortages? *Semiconductor Engineering* (Nov. 2021). https://semiengineering.com/end-in-sight-for-chip-shortages/

[66] Mark Lapedus. 2022. 200mm Shortages May Persist For Years. *Semiconductor Engineering* (Jan. 2022). https://semiengineering.com/200mm-shortages-may-persist-for-years/

[67] Alex Leary and Paul Zibro. 2021. Biden Calls for $50 Billion to Boost U.S. Chip Industry. *The Wall Street Journal* (March 2021). https://www.wsj.com/articles/biden-urges-50-billion-to-boost-chip-manufacturing-in-u-s-11617211570

[68] Yimou Lee and Ben Blanchard. 2021. TSMC announces chip plant in Japan, flags 'tight' capacity throughout 2022. *Reuters* (Oct. 2021). https://www.reuters.com/technology/taiwans-tsmc-posts-138-rise-q3-profit-global-chip-demand-surge-2021-10-14/

[69] Jenny Leonard. 2021. China's Auto-Chip Hoarding Probe Should Be Worrying Distributors. *Bloomberg* (Aug. 2021). https://www.bloomberg.com/news/articles/2021-09-23/white-house-pushes-companies-to-be-transparent-on-chips-supply

[70] Jenny Leonard. 2021. White House Pushes Companies to Be Transparent on Chips Supply. *Bloomberg* (Sept. 2021). https://www.bloomberg.com/news/articles/2021-09-23/white-house-pushes-companies-to-be-transparent-on-chips-supply

[71] Kif Leswing. 2021. GlobalFoundries CEO: We're sold out of semiconductor chip capacity through 2023. *CNBC* (Oct. 2021). https://www.cnbc.com/2021/10/30/globalfoundries-ceo-were-sold-out-of-semiconductor-chip-capacity-through-2023.html

[72] Ang Li and David Wentzlaff. 2021. PRGA: An Open-Source FPGA Research and Prototyping Framework. In *The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* (Virtual Event, USA) *(FPGA '21)*. Association for Computing Machinery, New York, NY, USA, 127–137. https://doi.org/10.1145/3431920.3439294

[73] Sean Lie. 2021. Multi-Million Core, Multi-Wafer AI Cluster. In *2021 IEEE Hot Chips 33 Symposium (HCS)*. IEEE Computer Society, 1–41.

[74] Paul Lienert. 2021. GM aims to tackle chip shortage with new designs made in North America. *Reuters* (Nov 2021). https://www.reuters.com/business/autos-transportation/gm-aims-tackle-chip-shortage-with-new-designs-made-north-america-2021-11-18/

[75] Junyi Lin, Virginia L.M. Spiegler, and M.M. Naim. 2018. Dynamic analysis and design of a semiconductor supply chain: a control engineering approach. *International Journal of Production Research* 56, 13 (2018), 4585–4611. https://doi.org/10.1080/00207543.2017.1396507 arXiv:https://doi.org/10.1080/00207543.2017.1396507

[76] Andy Longford. 2005. Chip packaging challenges … a roadmap based overview. *Microelectronics International* 22 (08 2005), 17–20. https://doi.org/10.1108/13565360510592180

[77] David MaCabe and Raymond Zhong. 2020. Trump Administration Widens Huawei Dragnet. *The New York Times* (Aug. 2020). https://www.nytimes.com/2020/08/17/technology/trump-huawei-commerce-chips.html

[78] Ikuo Magaki, Moein Khazraee, Luis Vega Gutierrez, and Michael Bedford Taylor. 2016. ASIC Clouds: Specializing the Datacenter. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*. 178–190. https://doi.org/10.1109/ISCA.2016.25

[79] Peter Milder, Franz Franchetti, James C. Hoe, and Markus Püschel. 2012. Computer Generation of Hardware for Linear Digital Signal Processing Transforms. *ACM Trans. Des. Autom. Electron. Syst.* 17, 2, Article 15 (apr 2012), 33 pages. https://doi.org/10.1145/2159542.2159547

[80] Takashi Mochizuki. 2021. It's Going to Get Even Harder to Buy a PlayStation 5. *Bloomberg* (Nov. 2021). https://www.bloomberg.com/news/articles/2021-11-11/sony-trims-playstation-5-assembly-plans-after-chip-shortages-hit

[81] Samuel Moore. 2021. How and When the Chip Shortage Will End, in 4 Charts. *IEEE Spectrum* (June 2021). https://spectrum.ieee.org/chip-shortage

[82] Samuel Moore. 2021. How to Keep the Automotive Chip Shortage From Happening Again. *IEEE Spectrum* (July 2021). https://spectrum.ieee.org/automotive-chip-shortage

[83] S.S. Mukherjee, C. Weaver, J. Emer, S.K. Reinhardt, and T. Austin. 2003. A systematic methodology to compute the architectural vulnerability factors for a high-performance microprocessor. In *Proceedings. 36th Annual IEEE/ACM International Symposium on Microarchitecture, 2003. MICRO-36.* 29–40. https://doi.org/10.1109/MICRO.2003.1253181

[84] Lars Mönch, Reha Uzsoy, and John W. Fowler. 2018. A survey of semiconductor supply chain models part I: semiconductor supply chains, strategic network design, and supply chain simulation. *International Journal of Production Research* 56, 13 (2018), 4524–4545. https://doi.org/10.1080/00207543.2017.1401233 arXiv:https://doi.org/10.1080/00207543.2017.1401233

[85] Samuel Naffziger, Noah Beck, Thomas Burd, Kevin Lepak, Gabriel H. Loh, Mahesh Subramony, and Sean White. 2021. Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. 57–70. https://doi.org/10.1109/ISCA52012.2021.00014

[86] Samuel Naffziger, Kevin Lepak, Milam Paraschou, and Mahesh Subramony. 2020. AMD Chiplet Architecture for High-Performance Server and Desktop Products. In *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*. 44–45. https://doi.org/10.1109/ISSCC19947.2020.9063103

[87] Board of Governors of the Federal Reserve System (USA). 2023. Industrial Production and Capacity Utilization. https://www.federalreserve.gov/releases/g17/

[88] Timothy Pettit, Joseph Fiksel, and Keely Croxton. 2010. Ensuring Supply Chain Resilience: Development of a Conceptual Framework. *Journal of Business Logistics* 31 (Mar 2010), 1 – 21. https://doi.org/10.1002/j.2158-1592.2010.tb00125.x

[89] Fred Y. Philips. 2001. *Market-Oriented Technology Management.* Springer.

[90] CS Premachandran, Thuy Tran-Quinn, Lloyd Burrell, and Patrick Justison. 2019. A Comprehensive Wafer Level Reliability Study on 65nm Silicon Interposer. In *2019 IEEE International Reliability Physics Symposium (IRPS)*. 1–8. https://doi.org/10.1109/IRPS.2019.8720515

[91] Noel Randewich. 2011. Thai floods, hard drive shortage threaten PC sales. *Reuters* (Oct 2011). https://www.reuters.com/article/us-thailand-floods-tech/thai-floods-hard-drive-shortage-threaten-pc-sales-idUSTRE79K76Z20111021

[92] Haoxing Ren, Kevin Bercaw, Tom Chadwick, Tom Guzowski, Juergen Koehl, Jeff Miller, and Steven Urish. 2007. How to process a multi million gate ASIC layout in 21 hours. In *2007 7th International Conference on ASIC*. 1118–1121. https://doi.org/10.1109/ICASIC.2007.4415829

[93] Renesas. 2021. UPDATE 10 - Notice Regarding the Semiconductor Manufacturing Factory (Naka Factory) Fire: Production Capacity Recovery Status. https://www.renesas.com/us/en/about/press-room/update-10-notice-regarding-semiconductor-manufacturing-factory-naka-factory-fire-production-capacity

[94] Reuters. 2021. BE Semiconductor cuts Q4 revenue guidance due to Malaysia floods. *Reuters* (Dec 2021). https://www.reuters.com/markets/europe/be-semiconductor-cuts-q4-revenue-guidance-due-flooding-malaysia-2021-12-20/

[95] Reuters. 2022. Chip undersupply to last until 2024, says Volkswagen CFO. *Reuters* (April 2022). https://www.reuters.com/business/autos-transportation/chip-undersupply-last-until-2024-says-volkswagen-cfo-boersen-zeitung-2022-04-09/

[96] Reuters. 2022. Smartphone shipments within China down 31.8% year-on-year in February, government data shows. *Reuters* (March 2022). https://www.reuters.com/world/china/smartphone-shipments-within-china-down-318-year-on-year-feb-govt-data-2022-03-21/

[97] Joshua J. Romero. 2012. The Lessons of Thailand's Flood. *IEEE Spectrum* (Nov 2012). https://spectrum.ieee.org/the-lessons-of-thailands-flood

[98] The White House Briefing Room. 2022. FACT SHEET: Joined by Allies and Partners, the United States Imposes Devastating Costs on Russia. https://www.whitehouse.gov/briefing-room/statements-releases/2022/02/24/fact-sheet-joined-by-allies-and-partners-the-united-states-imposes-devastating-costs-on-russia/

[99] Tim Ryan. 2022. Chips and Science Act. https://www.commerce.senate.gov/2022/7/view-the-chips-legislation

[100] Samsung. 2022. Samsung Exynos. https://www.samsung.com/semiconductor/minisite/exynos/

[101] Yakun Sophia Shao, Brandon Reagen, Gu-Yeon Wei, and David Brooks. 2014. Aladdin: A Pre-RTL, Power-Performance Accelerator Simulator Enabling Large Design Space Exploration of Customized Architectures. In *Proceeding of the 41st Annual International Symposium on Computer Architecuture* (Minneapolis, Minnesota, USA) *(ISCA '14)*. IEEE Press, 97–108.

[102] Sam Shead. 2021. Why Intel and TSMC are building water-dependent chip factories in one of the driest U.S. states. *CNBC* (Jun 2021). https://www.cnbc.com/2021/06/04/why-intel-tsmc-are-building-water-dependent-chip-plants-in-arizona.html

[103] Anton Shilov. 2007. IBM Produces Cell Processor Using New Fabrication Technology. *Xbit Laboratories* (March 2007).

[104] SiFive. 2021. SiFive and Samsung Foundry Extend Partnership to Accelerate AI SoC Development. https://www.sifive.com/press/sifive-and-samsung-foundry-extend-partnership-to-accelerate

[105] Teja Singh, Sundar Rangarajan, Deepesh John, Russell Schreiber, Spence Oliver, Rajit Seahra, and Alex Schaefer. 2020. 2.1 Zen 2: The AMD 7nm Energy-Efficient High-Performance x86-64 Microprocessor Core. In *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*. 42–44. https://doi.org/10.1109/ISSCC19947.2020.9063113

[106] SMIC. 2022. SMIC Backend Service. https://www.smics.com/en/site/posterior_segment

[107] Ilya M. Sobol. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation* 55, 1 (2001), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6 The Second IMACS Seminar on Monte Carlo Methods.

[108] Ed Sperling. 2018. Design Rule Complexity Rising. *Semiconductor Engineering* (April 2018). https://semiengineering.com/design-rule-complexity-rising/

[109] STMicroelectronics. [n. d.]. Arm Cortex-M in a nutshell. https://www.st.com/content/st_com/en/arm-32-bit-microcontrollers.html

[110] Jon Stokes. 2008. IBM shrinks Cell to 45nm. Cheaper PS3s will follow. *Ars Technica* (Feb. 2008). https://arstechnica.com/gaming/2008/02/ibm-shrinks-cell-to-45nm-cheaper-ps3s-will-follow/

[111] Dylan Stow, Yuan Xie, Taniya Siddiqua, and Gabriel H. Loh. 2017. Cost-effective design of scalable high-performance systems using active and passive interposers. In *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. 728–735. https://doi.org/10.1109/ICCAD.2017.8203849

[112] Ana Swanson and Raymond Zhong. 2020. U.S. Places Restrictions on China's Leading Chip Maker. *The New York Times* (Sept. 2020). https://www.nytimes.com/2020/09/26/technology/trump-china-smic-blacklist.html

[113] Synopsys. 2022. Synopsys Standard Cell Libraries. https://www.synopsys.com/dw/ipdir.php?ds=dwc_standard_cell

[114] Insignis Tech. 2018. The Advantages of Legacy Process Nodes. https://insignis-tech.com/technology/market-conditions/advantages-legacy-process-nodes/

[115] TechPowerUp. 2022. AMD "Matisse" and "Rome" IO Controller Dies Mapped Out. (Apr 2022). https://www.techpowerup.com/266287/amd-matisse-and-rome-io-controller-dies-mapped-out

[116] TSMC. 2020. TSMC 2020 Annual Report. https://investor.tsmc.com/static/annualReports/2020/english/index.html/

[117] TSMC. 2022. TSMC Advanced Packaging. https://www.tsmc.com/english/dedicatedFoundry/services/advanced-packaging

[118] TSMC. 2022. TSMC Financial Results. https://investor.tsmc.com/english/quarterly-results/

[119] TSMC. 2022. TSMC Logic Technology. https://www.tsmc.com/english/dedicatedFoundry/technology/logic

[120] Benjamin Tukamuhabwa, Mark Stevenson, Jerry Busby, and Marta Zorzini Bell. 2015. Supply chain resilience: Definition, review and theoretical foundations for further study. *International Journal of Production Research* 53 (04 2015), 1–32. https://doi.org/10.1080/00207543.2015.1037934

[121] UMC. 2022. UMC Turnkey Solutions. https://www.umc.com/en/StaticPage/turnkey_solutions

[122] Ganesh Venkatesh, Jack Sampson, Nathan Goulding, Saturnino Garcia, Vladyslav Bryksin, Jose Lugo-Martinez, Steven Swanson, and Michael Bedford Taylor. 2010. Conservation Cores: Reducing the Energy of Mature Computations. In *Proceedings of the Fifteenth International Conference on Architectural Support for Programming Languages and Operating Systems* (Pittsburgh, Pennsylvania, USA) *(ASPLOS XV)*. Association for Computing Machinery, New York, NY, USA, 205–218. https://doi.org/10.1145/1736020.1736044

[123] Ursula von der Leyen. 2022. Statement by President von der Leyen on the European Chips Act. https://ec.europa.eu/commission/presscorner/detail/en/statement_22_866

[124] Michael Wayland. 2021. Chip shortage expected to cost auto industry $210 billion in revenue in 2021. *CNBC* (Oct. 2021). https://www.cnbc.com/2021/09/23/chip-shortage-expected-to-cost-auto-industry-210-billion-in-2021.html

[125] Jeanne Whalen and Aaron Gregg. 2021. Seeking more reliable supply, Ford signs a deal with a huge chip maker. *The Washington Post* (Nov 2021). https://www.washingtonpost.com/business/2021/11/18/ford-computer-chip-globalfoundries/

[126] Ron Wyden. 2021. S.2107 - FABS Act. https://www.congress.gov/bill/117th-congress/senate-bill/2107

[127] Stephanie Yang. 2021. The Chip Shortage Is Bad. Taiwan's Drought Threatens to Make It Worse. *The Wall Street Journal* (April 2021). https://www.wsj.com/articles/the-chip-shortage-is-bad-taiwans-drought-threatens-to-make-it-worse-11618565400

[128] 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/

[129] F. Zaruba and L. Benini. 2019. The Cost of Application-Class Processing: Energy and Performance Analysis of a Linux-Ready 1.7-GHz 64-Bit RISC-V Core in 22-nm FDSOI Technology. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 27, 11 (Nov 2019), 2629–2640. https://doi.org/10.1109/TVLSI.2019.2926114

[130] Marcela Zuluaga, Peter Milder, and Markus Püschel. 2016. Streaming Sorting Networks. *ACM Trans. Des. Autom. Electron. Syst.* 21, 4, Article 55 (may 2016), 30 pages. https://doi.org/10.1145/2854150