Acceleration Potential in the Chip Design-To-Manufacturing Pipeline

Anonymous Authors[1]

Abstract

I analyze the process steps for a GPU New Product Introduction (NPI), i.e. the process for getting a new chip design from tapeout to mass production, focusing on existing technology nodes. The speed of NPI is important, as it determines the strength of the compute-compute feedback loop, which in turn influences the speed of AI development. I gives a qualitative overview of the NPI process, with detailed descriptions of individual process steps, present time estimates for individual NPI steps, and list potential levers for acceleration. I find that a further acceleration of the NPI process would likely be possible if design firms are willing to spend more resources on the process than currently economically viable.

1. Introduction

Recent progress in AI capabilities was driven to a significant extent by scaling of computational resources. As AI systems become more capable, and able to automate more complex industrial processes, they may become increasingly helpful for generating more compute. This potential for a compute - AI feedback loop - where more powerful compute enables more capable AI, which in turn designs even more powerful compute - could accelerate the speed of AI development. The velocity of this feedback loop, however, depends on the efficiency of the New Product Introduction (NPI) cycle: the timeline from a finalized chip design to its availability in mass production.

This report analyzes the NPI process for GPUs manufactured within existing, established semiconductor process nodes, excluding the development of new nodes themselves. I aim to map the critical process steps, quantify their typical durations, identify key bottlenecks, and explore potential strategies for acceleration. The findings are derived from a comprehensive review of technical literature and from in-depth interviews with 15 senior engineers and managers across various leading chip design and manufacturing firms. While this analysis focuses on current technological paradigms, the granular process mapping it provides could serve as a valuable baseline for future investigations into AI-driven acceleration of chip production. The full detailed analysis, including citations, is available in the appendix.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. The GPU New Product Introduction (NPI) Process

The NPI process is a sequence of intricate, often interdependent, stages transforming a digital GPU design into a mass-produced physical product.

2.1. Tapeout

Tapeout marks the formal conclusion of the chip design phase, at which the logical and physical architecture of the GPU is finalized and has undergone extensive pre-silicon verification. This includes simulations, virtual hardware emulations, and rigorous design rule checks (DRCs) against the foundry's Process Design Kit (PDK) to ensure manufacturability. Iterative communication between the design firm and the foundry is common leading up to tapeout, to resolve potential manufacturing issues.

Beyond electrical design, several non-electrical features are incorporated into the chip layout. These include company logos, copyright information, chip identifiers, and revision markers. Crucially, Design-for-Manufacturability (DfM) structures like protective seal rings around the die and filler structures to ensure planarity during chemical-mechanical polishing (CMP) are added. The final design is then arranged into a reticle layout, which involves creating a matrix of multiple chip copies (dies), adding scribe lines for die separation, integrating test patterns for in-process monitoring, and placing alignment marks for lithographic precision. Once complete, the design data (e.g., in GDSII or OASIS format) is transferred, often via secure File Transfer Protocol (FTP), to a mask shop. The project management activities surrounding tapeout, including database management, data entry, and various sign-offs, can consume up to two weeks. It's important to note that a single product might undergo multiple tapeouts if design flaws are discovered during ini-

tial prototyping, leading to iterated designs.

2.2. Photomask Production

Upon receiving the design data, the mask shop - either a captive facility owned by the foundry (e.g., TSMC, Intel, Samsung) or an independent merchant shop (e.g., Dai Nippon Printing) - takes on the production of photomasks. This stage is critical as masks are the master templates for patterning wafers. It can be broadly divided into computational data preparation and physical manufacturing.

2.2.1. Post-Tapeout Flow (PTOF)

PTOF comprises computationally intensive steps to prepare the design data for the mask writing tools.

- **Resolution Enhancement Techniques (RET):** As chip features shrink below the wavelength of lithography light, RET becomes essential to correct for optical diffraction effects. Optical Proximity Correction (OPC) involves adding sub-resolution assist features or modifying pattern shapes on the mask. Inverse Lithography Technology (ILT) is a more advanced, computationally demanding form that reverse-engineers the optimal mask pattern from the desired wafer pattern, creating complex curvilinear mask shapes. While historically slow (ILT TAT could be days to weeks), the advent of GPU-accelerated computational lithography (e.g., NVIDIA's CuLitho platform) and multi-beam mask writers has dramatically reduced RET processing times.

- **Mask Process Correction (MPC):** MPC compensates for systematic errors in pattern critical dimensions (CD) that arise during the mask manufacturing process itself, such as e-beam proximity effects, resist development, and etching biases. This involves adjusting mask pattern shapes and locally varying e-beam exposure doses. MPC is particularly important for advanced nodes ($<$16nm) and EUV lithography. The complexity of ILT-generated patterns can significantly increase MPC runtime, but techniques like pixel-level dose correction (PLDC) aim to integrate MPC into the mask writing step. MPC TAT can range from hours to days, depending on compute resources and pattern complexity.

- **Mask Data Preparation (MDP) / Fracturing:** This step converts the corrected, complex design layout (often terabytes per layer for modern chips due to curvilinear features and high vertex counts) into a simpler, machine-readable format (e.g., MEBES, OASIS.MASK) that the mask writer can process. File sizes have exploded with EUV and ILT adoption. MDP TAT typically ranges from a few hours to over 17 hours for the most complex masks.

2.2.2. Physical Mask Manufacturing, Inspection, and Repair

Once PTOF is complete, physical mask fabrication begins.

- **Mask Writing and Pattern Transfer:** A mask writer exposes a pattern onto a mask blank—a quartz substrate coated with an absorber material (e.g., chrome for DUV, tantalum-based for EUV) and photoresist. EUV blanks are complex multi-layer structures. Laser writers are used for less critical layers, while electron-beam (e-beam) writers are used for critical layers. Variable Shape Beam (VSB) e-beam writers, which expose patterns sequentially, face throughput challenges with increasing complexity. Multi-Beam Mask Writers (MBMs) overcome this by using thousands of parallel beams, enabling constant write times (typically 10-12 hours per mask) regardless of pattern complexity. After exposure, the resist is developed, and the pattern is etched into the absorber layer.

- **Mask Inspection, Metrology, and Repair:** Masks undergo stringent quality control. Inspection tools (optical for larger features, e-beam for high resolution, and actinic for EUV wavelength defect review, e.g., Zeiss AIMS EUV) detect defects like particles (soft defects) or pattern errors (hard defects). Metrology tools (e.g., SEM, AFM) measure CDs, pattern placement (registration), and overlay accuracy. If defects are found, repair is attempted using tools like focused ion beam (FIB), e-beam, or nanomachining. Multiple inspection loops (pre- and post-pellicle attachment) are common. This entire sequence can take hours to days per mask. Despite increasing defect challenges with smaller nodes, EUV mask yields were reported around 91% in 2020, with only a small fraction requiring return from the fab.

The average TAT for a critical-layer mask (7-11nm) was reported as $\sim$7.53 days in a 2020 industry survey, with a typical mask set comprising $\sim$76 masks, of which 20-30% are critical.

2.3. Prototyping

This phase is often the longest and most uncertain in the NPI cycle.

1. **Engineering Sample (ES) Manufacturing:** The first physical chips ("first silicon") are produced using the newly created mask set. This involves running a small number of wafer lots through the full fab process. Design firms may use Multi-Project Wafer (MPW) ser-

vices to share mask and wafer costs for initial prototypes. Fabs prepare manufacturing lines with pilot lots ("pipe-cleaners") and may run corner lots to test process window variations. The manufacturing itself involves hundreds of steps (deposition, lithography, etch, implant, CMP, etc.). The TAT for ES manufacturing is primarily driven by the number of mask layers and the fab's cycle time per layer (Days Per Mask Layer - DPML), typically ranging from 2 to 5 months for cutting-edge GPUs. Effective Design-Technology Co-Optimization (DTCO) is crucial for aligning design with manufacturing capabilities.

2. **Post-Silicon Validation (PSV):** Once ES wafers are diced, packaged, and assembled onto test boards, they undergo extensive PSV. This is a critical debugging stage to ensure the chip functions as intended and meets performance, power, and reliability targets. PSV includes power-on debug, basic hardware logic validation, hardware/software co-validation, electrical validation (I/O, power delivery, clocking), and speed-path analysis . Unlike pre-silicon simulation, PSV tests the actual silicon at speed but suffers from limited internal observability and controllability, making debug challenging. PSV is a major time commitment, often lasting 3 to 9 months, consuming significant engineering resources and constituting a large portion of overall design costs.

3. **Respins:** If PSV uncovers critical flaws that cannot be fixed through workarounds or minor process adjustments, a respin is necessary. This involves iterating the chip design, creating a new set of (at least some) photomasks, manufacturing new engineering samples, and re-validating. Respins are costly in both time and resources. First-silicon success is not the norm; a 2022 study indicated only ∼25% of IC/ASIC projects required no respins, with logic defects being the most common cause. Each respin can add 1 to 5 months to the NPI timeline, depending on the nature and location of the defect in the chip's layer stack.

2.4. High-Volume Manufacturing (HVM) Ramp

After a chip design successfully passes all product release qualifications, it transitions to HVM. The focus shifts from debugging to process optimization for yield, throughput, and cost. Yields typically improve from ES levels (e.g., 90-95%) to mature HVM levels (e.g., 95-97% or higher). The ramp-up to full HVM efficiency can take several weeks to months (commonly 1.5-3 months for the direct manufacturing ramp). Some expert estimates extend this to 6-12 months if extensive system-level testing or lead customer validation is included as part of the HVM ramp qualification. Some companies engage in "risk production," starting HVM ramp-up before PSV is fully complete to accelerate time-to-market, though this carries the risk of shipping chips with undiscovered defects. Even with successful qualification, HVM can encounter new, volume-dependent yield issues, as exemplified by challenges in Intel's 10nm process ramp due to interconnect problems not apparent in smaller ES volumes.

3. NPI Duration Estimates and Key Bottlenecks

Aggregating insights from 15 industry experts, the NPI cycle time for cutting-edge GPUs (within existing nodes and assuming no respins) typically ranges from **approximately 7 to 15.8 months**. The mean lower bound across expert estimates is ∼8.5 months, and the mean upper bound is ∼13.6 months. Table 3 provides a summary of these estimates.

Process Step	Expert-Aggregated Duration (Months)
Photomask Production (until first wafers start)	0.1 – 1.5
Engineering Sample Manufacturing	2.0 – 5.0
Post-Silicon Validation (PSV)	3.0 – 9.0
Respin (if required, per iteration)	1.0 – 5.0
Ramp to HVM (post-PSV/qualification)	1.5 – 12.0[1]
Total NPI (Tapeout to HVM, no respins)	**7.0 – 15.8**

Table 1. Estimated Durations for GPU NPI Process Steps (Months).

These figures are subject to considerable variability due to factors like chip complexity (e.g., die size, layer count, use of advanced packaging like CoWoS), specific technology node maturity, foundry practices, client requirements, and the priority assigned to the project. The most consistently cited time-intensive stages, and thus primary bottlenecks, are:

- **Engineering Sample Manufacturing** (∼2-5 months)
- **Post-Silicon Validation** (∼3-9 months)

While individual photomask creation is a multi-day process per mask, the ability to manufacture masks in parallel and to commence wafer fabrication before the entire mask set is complete means that mask production, while critical, does not usually dictate the overall NPI critical path length to the same extent as prototyping and validation.

4. Potential Levers for NPI Acceleration

Despite strong existing commercial incentives to minimize NPI time, experts agree that significant further acceleration is plausible if speed becomes the overriding priority, potentially driven by a well-resourced actor unconstrained by typical commercial trade-offs. Key strategies identified include:

- **Radically Accelerating Post-Silicon Validation (PSV):** This involves massive parallelization and resource increases. Examples include deploying substantially more compute resources to accelerate test execution (especially for random or exhaustive tests), significantly increasing the number of engineering samples to get statistically robust data faster (potentially requiring dedicated, high-throughput prototyping lines), and employing larger, highly skilled validation teams working in continuous shifts. Designing chips with enhanced Design-for-Testability (DfT) features, such as modularity for easier defect isolation or increased internal observability, is crucial. Furthermore, "shifting left" as much validation as possible to pre-silicon stages through more extensive and faster emulation and simulation can reduce the PSV burden. Some experts even suggested having multiple independent PSV teams work in parallel on the same design.
- **Proactively Avoiding and Mitigating Respins:** Improving the anticipation of potential defects through advanced design tools that predict process variations and timing issues is key. This could involve manufacturing multiple mask variants for the most critical layers to have alternatives ready. Allowing greater flexibility in final product specifications can also prevent respins triggered by minor deviations from initial targets; custom silicon providers often have more leeway here than those serving broad markets with fixed feature sets.
- **Strategic Trade-offs in Quality and Yield for Speed:** A determined actor could accept significantly lower initial HVM yields (e.g., 10-20% instead of 90%+) to begin mass production much earlier. This might involve multiple "risk production" cycles, discarding many non-functional chips but gaining time. Similarly, if the end-use case is limited (e.g., a single, year-long training run), long-term reliability and durability testing (like High-Temperature Operating Life - HTOL tests) could be drastically reduced or eliminated, saving weeks or months. Reducing the number or stringency of "non-valuable" (for the specific urgent use-case) inspection and metrology steps throughout the manufacturing process could also save time, albeit at the risk of lower overall quality.
- **Optimizing Manufacturing Throughput (Wafers and Masks):**
  - *Wafer Fabrication:* All lots (ES and HVM) could be run at the absolute highest priority (e.g., "ultra super hot lot" status), minimizing DPML. Using smaller lot sizes can also increase agility. Fabs could be designed or reconfigured to eliminate inter-tool wait times, reducing cycle time to raw processing time. Increasing the capacity of bottleneck tool groups (e.g., EUV scanners) and meticulously identifying and mitigating "dog tools" (underperforming individual machines) are also vital.
  - *Mask Manufacturing:* Securing dedicated, high-priority mask production lines, free from competition for capacity, would minimize mask TAT. Continued investment in faster PTOF algorithms and compute infrastructure is needed to ensure recent speed gains persist with increasing chip complexity. Ensuring MBM writers and advanced inspection tools are fully leveraged and potentially run at lower overall fab utilization to prioritize specific mask sets can also contribute.
- **Holistic Process Re-engineering and AI Integration:** Designing the entire NPI process with speed as the primary design criterion from the outset, rather than optimizing existing sequential flows, could unlock further gains. This includes tight co-optimization of design, DfT, manufacturing processes, and validation strategies. Many experts anticipate that AI tools could play a significant role in various aspects of acceleration, from optimizing PTOF algorithms and fab scheduling to aiding in PSV debug and predicting manufacturing variances, though specific, quantified time savings from AI are still an area of active research and development.

5. Implications

The New Product Introduction cycle for state-of-the-art GPUs is a complex process, typically spanning 7 to 15.8 months even before accounting for potential design respins. The most substantial portions of this timeline are the manufacturing of engineering samples and the exhaustive post-silicon validation required to ensure functionality and performance. While the semiconductor industry operates under intense pressure to shorten these cycles, our analysis suggests that considerable further acceleration is possible. This would require a paradigm shift where speed is prioritized above traditional commercial considerations like cost per chip or maximizing yield on the first iteration, backed by exceptionally large resource commitments. As a result, new GPU designs could be brought into production faster, accelerating Moore's law, and hence speeding up AI development. To arrive at a more thorough understanding of the compute-compute feedback loop, more research is required on the acceleration potential of AI capabilities themselves on speeding up the NPI process.

Impact Statement

This paper presents work whose goal is to advance the field of technical AI governance. There are many potential soci-

etal consequences of our work, none which we feel must be specifically highlighted here.

References

Balasinski, A. Mask cycle time reduction for foundry projects. In *Photomask Technology 2011*, volume 8166 of *Proc. SPIE*, pp. 81660K, September 2011. doi: 10.1117/12.896736. URL https://doi.org/10.1117/12.896736. Alternative URL from footnote 14: https://www.sciencedirect.com/science/article/abs/pii/S073658451000244li.

Bork, I., Buck, P., Bürgel, C., Durvasula, B., Eder-Kapl, S., Hudek, P., Jurkovic, M., Klikovits, J., Platzgummer, E., Rankin, J. H., Nageswara, R., Reddy, M., and Spengler, C. Mask process correction validation for multi-beam mask lithography. In *Photomask Technology 2018*, volume 10810 of *Proc. SPIE*, pp. 108100K, November 2018. doi: 10.1117/12.2503284. URL https://doi.org/10.1117/12.2503284.

Egodage, K., Tu, F., Schneider, H., Hermanns, C. F., Kersteen, G., Szafranek, B., and Schulz, K. SEM autoanalysis for reduced turnaround time and to ensure repair quality of EUV photomasks. In *International Conference on Extreme Ultraviolet Lithography 2019*, volume 11147 of *Proc. SPIE*, pp. 111471G, September 2019. doi: 10.1117/12.2538474. URL https://doi.org/10.1117/12.2538474.

Gilgenkrantz, P., Riad, B., Salah, M., and Siddique, A. Cloud flight plan for post-tapeout flow jobs. In *DTCO and Computational Patterning III*, volume 12954 of *Proc. SPIE*, pp. 129540O, April 2024. doi: 10.1117/12.3010142. URL https://doi.org/10.1117/12.3010142.

Maniyara, R. A., Ghosh, D. S., and Pruneri, V. Transparent and conductive backside coating of EUV lithography masks for ultra short pulse laser correction. In *Photomask Technology 2017*, volume 10451 of *Proc. SPIE*, pp. 104511S, October 2017. doi: 10.1117/12.2280526. URL https://doi.org/10.1117/12.2280526.

Matsumoto, H., Yasuda, J., Motosugi, T., Kimura, H., Kojima, Y., Yamashita, H., Saito, M., and Nakayamada, N. Multi-beam mask writer MBM-3000 for next generation EUV mask production. In *Novel Patterning Technologies 2024*, volume 12956 of *Proc. SPIE*, pp. 1295602, April 2024. doi: 10.1117/12.3010686. URL https://doi.org/10.1117/12.3010686.

Mishra, P., Morad, R., Ziv, A., and Ray, S. Post-silicon validation in the SoC era: A tutorial introduction. *IEEE Design & Test*, 34(3):68–92, June 2017. doi: 10.1109/MDAT.2017.2691348. URL https://doi.org/10.1109/MDAT.2017.2691348.

Nahir, A. et al. Bridging pre-silicon verification and post-silicon validation. In *Design Automation Conference*, pp. 94–95, Anaheim, CA, USA, 2010. doi: 10.1145/1837274.1837300. URL https://doi.org/10.1145/1837274.1837300.

NuFlare Technology. MBM-2000 Series. https://www.nuflare.co.jp/english/products/beam/mbm_2000/, Accessed on 2025-05-12. URL from footnote 71. Actual publication date not specified.

Pang, L., Vidal-Russell, E., Baggenstoss, B., Lee, M., Digaum, J. L., Yang, M.-C., Ungar, P. J., Bouaricha, A., Wang, K., Su, B., Pearman, R., and Fujimura, A. Breakthrough curvilinear ILT enabled by multi-beam mask writing. *Journal of Micro/Nanopatterning, Materials, and Metrology*, 20(4):041405, November 2021. doi: 10.1117/1.JMM.20.4.041405. URL https://doi.org/10.1117/1.JMM.20.4.041405.

Schneider, H., Tu, F., Ahmels, L., Szafranek, B., Gries, K., Rhinow, D., Vollmar, S., Krugmann, A., Schoenberger, R., Pauls, W., Verch, A., Capelli, R., Vincenzo, A. D., Kersteen, G., Marbach, H., and Waldow, M. High-end EUV photomask repairs for 5nm technology and beyond. In *Photomask Technology 2020*, volume 11518 of *Proc. SPIE*, pp. 1151808, October 2020. doi: 10.1117/12.2572879. URL https://doi.org/10.1117/12.2572879.

Shin, K., Kim, S., Choi, J., Kim, E., Lee, S., Hoeller-Lugmayr, H., Moqanaki, A., Schnettelker, A., Tomandl, M., Zillner, C., Torigoe, Y., Fujii, T., Miyake, R., Odaka, M., Masumoto, I., and Hamaji, M. New multi-beam mask data preparationmethod for EUV high volume data. In *Photomask Japan 2023: XXIX Symposium on Photomask and Next-Generation Lithography Mask Technology*, volume 12915 of *Proc. SPIE*, pp. 129150D, September 2023. doi: 10.1117/12.2683168. URL https://doi.org/10.1117/12.2683168.

Sugimori, T., Ogawa, R., Takekoshi, H., Hartley, J. G., Pinkney, D. J., Ando, A., Ishii, K., Noda, C., and Kikuiri, N. Study of high throughput EUV mask pattern inspection technologies using multi e-beam optics. In *Photomask Technology 2021*, volume 11855 of *Proc. SPIE*, pp. 118550C, September 2021. doi: 10.1117/12.2600987. URL https://doi.org/10.1117/12.2600987.

Tsunoda, D., Torigoe, Y., Sato, Y., Hamaji, M., Chua, G.-S., and Bürgel, C. Applying MPC for EUV mask fabrication. In *Photomask Technology 2018*, volume 10810 of *Proc.*

SPIE, pp. 108101H, October 2018. doi: 10.1117/12.2502068. URL https://doi.org/10.1117/12.2502068.

Tsunoda, D., Horima, Y., Torigoe, Y., Dillon, B., Lyons, A., Wallow, T., Wampler, K., Shu, V., and Zhang, Q. Free form data reduction for MPC. In *Photomask Technology 2020*, volume 11518 of *Proc. SPIE*, pp. 115180P, September 2020. doi: 10.1117/12.2573132. URL https://doi.org/10.1117/12.2573132.

Zillner, C., Moqanaki, A., Höller, H., Platzgummer, E., Hamaji, M., and Fujii, T. File formats for curvilinear multi-beam writing of 193i and EUV masks. In *Photomask Technology 2021*, volume 11855 of *Proc. SPIE*, pp. 1185512, October 2021. doi: 10.1117/12.2602375. URL https://doi.org/10.1117/12.2602375.

Ziv, A. Post-silicon validation (slides). https://theory.stanford.edu/~barrett/fmcad/slides/3_Ziv.pdf, Accessed on 2025-05-12. URL from footnote 190. Referenced in context of Ziv (2019). Actual presentation date not specified.

A. Details on process steps

A.1. Tapeout

Tapeout is commonly defined as the point at which a chip design is completed and handed over to the foundry for manufacturing. Within the development of a single chip there are often multiple tapeouts.[2] During the first phase of prototyping, it is common that design flaws are discovered, so that the chip design firm has to make slight changes to their design. When the iterated design leaves the firm, this is called a second tapeout (or third, or forth, and so on).

Before a chip is ready for tapeout, its electrical design features are finalized and have undergone extensive verification. For example, the logic design of a chip is tested using virtual hardware emulations[3], and there are design rule checks that ensure that the design adheres to the rules laid out by the foundry's process design kit (PDK). In many instances, this involves iterative processes between the design firm and the foundry: the foundry communicates which design elements might lead to manufacturing failures at their fab, design firm iterates based on feedback.[4]

Moreover, before a design is sent to the foundry, a number of non-electrical design features are added to the chip layout.[5] The company logo, copyright information, chip identification, and design revision information are added on the surface area. Design-for-manufacturability (DfM) structures are added, such as a protective seal ring, and filler structures to improve planarity. Multiple copies of the chip are arranged in a matrix format to create the reticle layout. This also involves adding scribe lines to separate dies, test patterns for process monitoring, alignment marks to align masks during exposure, and markers for precise laser cutting or sawing.

Once these steps are completed, the file is sent to a mask shop using a File Transfer Protocol (FTP)[6][7], using a standardized file format.[8] Balasinski (2011) claims that the project management activities associated with a typical tapeout can take up to two weeks, which includes handling of the design database, manual data entry, as well as various steps involving sign-offs and approvals.[9] (Balasinski, 2011)

A.2. Photomask production

Once the chip design is transferred, the mask shop is in charge of preparing and manufacturing the photomasks. Large firms like TSMC, Samsung, and Intel are known to have their own mask shops (captive mask shops).[10] There are also independent (merchant) mask shops, such as Dai Nippon. In 2013, the market was split approximately half-half between captive mask shops and merchant mask shops.[11]

The most time-consuming steps for mask production can roughly be divided into two clusters of processes that take place in chronological order:[12]

- Post-tapeout flow:[13] this involves computational processes aimed at preparing the chip design data to be processed by

[2] https://www.eetimes.com/vendors-should-count-silicon-not-tapeout-wins/

[3] Expert 7

[4] Expert 6, Expert 7

[5] For a more detailed definition of tapeout, see Lienig and Scheible (2020) p.103.

[6] https://www.sciencedirect.com/science/article/abs/pii/S0736584510000244li

[7] Expert 8 claims that it takes around 20 hours to transfer a design file, but it is unclear how representative this is, as it depends on the file size and performance of the FTP.

[8] For the latest chip generations, this is usually OASIS (Open Artwork System Interchange Standard), another well-known format is GDSII.

[9] It is plausible that this time span is not accurate anymore, as many steps may have been automated during the last 13 years.

[10] https://www.tomshardware.com/pc-components/cpus/nvidias-generative-ai-tool-delivers-a-radical-60x-performance-boost-for-chipmakers-tsmc-and-synopsys-are-now-using-the-culitho-software-in-production

[11] https://en.wikipedia.org/wiki/Mask$_s hop$

[12] Terminology and exact subdivisions vary between authors. Yamabe et. al. (2008) use this subdivision to calculate mask turnaround time, but acknowledge that this is a simplification. Rivzi (2005) makes a similar subdivision, calling the three steps "Data Preparation", "Front End of Line" and "Back End of Line". Plausibly, there are subdivisions that differentiate between more than three steps. For example Hong et.al. (2023) include an additional verification step.

[13] Naming of this step is inconsistent across the literature. Liening et. al. (2020, p.103) call this "Layout-to-Mask Preparation". Yamabe (2008) and Rizvi (2005) refer to this step as (Mask) Data Preparation (MDP), but confusingly, MDP sometimes only refers to fracturing and at other times includes both fracturing and OPC. To avoid confusion I will refer to this step as "Post-tapeout flow" (PTOF) in line with Salah (2023) at Siemens.

mask writers. Important process steps in this cluster are resolution enhancement techniques, mask process correction, and mask data preparation.

- Mask manufacturing, inspection, and repair: this involves physical manufacturing and testing of photomasks. Important process steps in this cluster are e-beam lithography, pattern transfer, mask inspection and repair, as well as cleaning and pellicalization.[14]

The time between a mask design arriving at the mask shop and the point of mask delivery can be a substantial contributor to overall cycle time in the design-to-manufacturing lifecycle. In the photomask industry, this span of time is called turnaround time or TAT. I will be using this term throughout the analysis.[15]

The latest available data on photomask TAT comes from an industry survey from 2020.[16] It states that the average TAT for nodes between 7nm and 11nm was 7.53 days per critical-layer mask. The 2019 version of this survey states that in the 11nm–7nm category, an average mask set had 76 masks. Chua et al. (2011) claims that around 20–30% of these layers are critical layers, although it is unclear how this has developed since 2011.

There are a few pieces of evidence that seem to contradict these estimates. For example, an expert from GlobalFoundries stated that they "could get one mask every 20–22 hours for 7nm"[17], but it is unclear if he was referring to critical or non-critical layers. A news article from 2022[18] mentions that average delivery time for "high-spec (photomask) products" has increased from 7 days to 30–50 days, but it is also unclear what kinds of masks they refer to, if they even refer to photomasks at all, and whether this is a temporary phenomenon. However, it seems that the data from the industry survey is more reliable than these anecdotal bits, as it aggregates input from 10 important companies in the space.[19]

A.2.1. Post-tapeout flow (PTOF)

The chip layout coming out of the design firm is not yet ready for photomask production. The post-tapeout flow translates a chip layout from an input file format that is not readable by a mask writer (e.g. OASIS), into an output file format that is readable by a mask writer (e.g. MEBES, OASIS.MASK).

The PTOF can be significant for TAT because it involves a number of computationally intensive processing steps, such as inverse computational lithography, mask process correction, and mask data preparation. The complexity of these processing steps increases fast as critical dimensions shrink, but recent years have seen some technological breakthroughs that have significantly accelerated key parts of the process.

The post-tapeout flow involves three core steps:[20]

- Resolution enhancement techniques (RET): a set of computational processes aimed at correcting for optical distortion effects that impact the precision with which circuit patterns can be etched onto a photomask.
- Mask Process Correction (MPC): a technique that corrects for systematic errors in photomask pattern critical dimension (CD) stemming from electron beam proximity effects, as well as from characteristics of the etch and development processes.

[14] Attachment of a protective membrane (pellicle) on the mask.

[15] It is somewhat unclear how TAT is defined. Aki Fujimura of the E-Beam Initiative seems to state in a video that it is the time period between "RET-out" to mask completion. However, in the same video, as well as in another video, he discusses OPC/ILT as a major component of TAT. Upon request, Jan Willis from the E-Beam Initiative clarified that TAT refers to the turnaround time for a single mask, and includes "everything from order to delivery", at least for the E-Beam Initiative industry surveys.

[16] https://www.ebeam.org/docs/eBeam-Mask-Maker-Survey-2020.pdf

[17] Expert 8

[18] https://english.etnews.com/20221108200001

[19] AMTC, DNP, HOYA, Intel, Micron, Photronics (incl. PDMC), Samsung, SMIC, TMC, and Toppan

[20] There is no broadly agreed terminology for each process step, and different authors use different subdivisions. Based on a survey of the available literature, my best guess is that RET, MPC, and MDP are the key process steps. A few examples of different taxonomies in the literature: Chua (2011) mentions OPC, MPC, and MDP (fracturing). Vu (2018) mentions RET and MDP, but groups MPC and fracturing into MDP. They also mention preparatory steps (e.g. pre-OPC). Endo (2012) mentions RET, OPC, MRC, and MDP (among others). Choi (2021) mentions OPC/ILT, MPC, and MDP. Calibre mention an integrated OPC, MPC and MDP software. Semiwiki mentions MDP and OPC/ILT, but not MPC. Bork and Buck (2018) mention OPC/MPC/Fracture/Write. Siemens mentions "SVRF" as an additional process step next to OPC, MPC, and MDP in one paper, but doesn't mention this in another paper. Another Semiwiki article mentions MEC, Pattern matching, MRC, and Fracture

- Mask Data Preparation (MDP): converting the OASIS data into a format readable by an e-beam writer.

Resolution Enhancement Technologies (RET)

RET is a collective term for techniques that modify photomasks in order to correct for diffraction effects.[21] This became necessary by the early 2000s[22] a few years after chip feature sizes became smaller than the wavelength of the lithography tools.[23] The most computationally intensive[24] forms of RET involve changing the shape of the photomask pattern in order to correct for irregularities in line width and spacing that may occur due to diffraction effects. For example, NVIDIA claims that the RET-related computational processes required for one mask set can exceed 30 million CPU hours, necessitating large data centers within foundries.[25] A basic form of RET is known as optical proximity correction (OPC). This involves adding unintuitive-looking, box-shaped features to the photomask pattern[26], which ensures that the printed pattern comes out in the shape desired by the designers.[27] Over time, a more sophisticated form of OPC emerged, which starts directly with the desired pattern on the wafer and then reverse-engineers how the photomask must look like, in order to achieve this pattern. This is known as inverse lithography technology (ILT[28]).

A TSMC expert suggests that, for an unspecified TSMC-manufactured chip, ILT can take several days to a week and OPC ∼1–3 days. By contrast, an Intel expert reports that OPC-type activities for an Intel Arc 5nm graphics chip took ∼1–2 weeks.[29] In general, ILT results in more complex, rounded (curvilinear) mask patterns, that are harder to calculate and write than OPC-style box-shaped designs.[30] Dillinger (2022) mentions that ILT computational runtime has been 20x slower than OPC and that single-beam VSB mask writers took days to expose an ILT mask[31] However, both of these bottlenecks were recently addressed by the introduction of multi-beam mask-writers which enable writing times that are constant, independently of mask complexity.[32] This speed-up has enabled the adoption of curvilinear masks, which had previously been impractical to manufacture.[33][34] On the computational side, there have also been advances. In 2023, NVIDIA introduced CuLitho, a library for GPU-accelerated computational lithography, that it claims can accelerate ILT by 40x.[35] NVIDIA CEO Jensen Huang further claimed that CuLitho will accelerate ILT time [37]from 2 weeks to 8 hours, and that TSMC can do their calculations on 350 DGX H100 GPUs instead of 40,000 CPUs[38][39] These advances suggest that ILT is currently not a bottleneck at the current margin. However, future increases in chip complexity may change the picture here, and it is unclear if the efficiency advantage of ILT will persist. In 2024, NVIDIA expected a 10x increase in computational load for future photomasks, although it is unclear which precise types of masks they are referring to.[40]

Mask Process Correction (MPC)

[21]The patterns for integrated circuits are projected on the silicon surface. As the light passes through the mask, diffraction occurs, so that the light is increasingly unfocused once it hits the surface. RET attempts to offset this unwanted effect by creating masks that correct for this diffraction. Creating these masks correctly is a difficult computational challenge.

[22]https://semiengineering.com/the-quest-for-curvilinear-photomasks/

[23]Also called sub-wavelength era.

[24]Additional RET techniques are phase shift masks (PSM), source mask optimization (SMO), and off-axis-illumination (OAI). There are also steps that are complementary to RET, such as mask rule checks (MRC), which aim to ensure that the mask adheres to a set of design rules.

[25]https://nvidianews.nvidia.com/news/tsmc-synopsys-nvidia-culitho

[26]For example, dog ears and sub-resolution assist features (SRAF).

[27]Initial approaches to this were rule-based OPC, and model-based OPC.

[28]Techniques like OPC and ILT are sometimes summarized under the term computational lithography. It should be noted that ILT is not used for all mask layers, but only the most complex critical layers.

[29]Expert 1, Expert 4

[30]These are also called Manhattan shapes. These are also called Manhattan shapes.

[31]https://semiwiki.com/semiconductor-manufacturers/tsmc/313540-inverse-lithography-technology-a-status-update/

[32]https://semiwiki.com/semiconductor-manufacturers/tsmc/313540-inverse-lithography-technology-a-status-update/

[33]https://semiengineering.com/the-quest-for-curvilinear-photomasks/

[34]There is a practical limit to photomask manufacturing. Mask write time needs to be limited to 24 hours or less to minimize overall defects, and for budgetary reasons, mask turnaround time (TAT) is optimized for 10 hours or less at the major mask shops. (Pearman et. al. (2019))

[35]https://spectrum.ieee.org/inverse-lithography[36]

[37]He refers to “mask design time” but my guess is that he is referring to ILT time.

[38]https://www.eetimes.com/nvidia-brings-gpu-acceleration-to-computational-lithography/

[39]https://www.computerbase.de/2024-03/nvidia-culitho-gpu-beschleunigte-lithografie-geht-in-die-serienproduktion/

[40]https://www.computerbase.de/2024-03/nvidia-culitho-gpu-beschleunigte-lithografie-geht-in-die-serienproduktion/

MPC is a technique that corrects for systematic errors in photomask pattern critical dimension (CD)[41] stemming from electron beam proximity effects,[42][43] as well as from characteristics of the etch and development processes.(Bork et al., 2018) MPC corrects these errors by adjusting the shape of the mask pattern, and by varying local e-beam exposure doses.[44] The technique is similar to RET in that they both aim to improve the accuracy of the patterns transferred onto the wafer. The difference is, roughly, that RET corrects for optical errors that occur during wafer exposure, while MPC corrects for errors that occur during photomask exposure.[45][46] MPC became especially prevalent for the tight CD specifications in nodes below 16nm,[47] and was described as a key enabling technology for process nodes beyond that.[48] (Bork et al., 2018) The introduction of special EUV mask substrates for nodes ¡7nm further reinforced the need for MPC. Survey data from 2017 shows that 72% of masks at below 7nm had MPC applied, with only ∼30–40% for ¡16nm/≥7nm, and ∼4% for ¡22nm/≥16nm. MPC was highlighted as a process step with potentially long runtime, and was called a potential bottleneck in the post-tapeout flow. This is not always justified unambiguously, but plausibly, it is the result of increasing pattern complexity associated with ILT. Tsunoda (2020) claims that ILT output has up to 15x more vertices than conventional OPC output, (Tsunoda et al., 2020) and according to Nakayamada (2022) MPC turnaround time goes up exponentially with pattern complexity. At the same time, there are approaches aiming to cut out MPC runtime entirely, by integrating MPC into the mask writing process through a technology called pixel-level dose correction (PLDC). NuFlare claim that their MBM-2000 mask writer applies this technology, but it is unclear to what extent this is adopted in industry, and if it has succeeded in cutting out MPC runtime. (NuFlare Technology, Accessed on 2025-05-12)

It is difficult to get precise industry-level runtime estimates for MPC. Partly this is due to the fact that speed depends on the amount of compute employed. For example, Gilgenkrantz (2024) provides Siemens Calibre runtimes between ∼13 hours and ∼1 hour for a 10nm layer, depending on the number of CPUs employed. (Gilgenkrantz et al., 2024) Tsunoda (2018) reports that MPC runtimes are approximately twice as long for EUV than for DUV. (Tsunoda et al., 2018) An expert at GlobalFoundries estimates that for a 12nm chip, MPC took ∼1–2 days.

Fracturing / mask data preparation (MDP)

Fracturing/MDP is the process of converting a mask design file consisting of complex geometrical figures into a set of simple shapes that are writable by a mask writer. Fracturing is a computationally intensive process due to the file sizes of post-RET chip design files and due to the increase in file size post-fracture. Already at the 28nm node, file sizes reached hundreds of gigabytes. In 2021, experts from Samsung and Siemens estimate that file sizes per mask layer are as much as several terabytes. In 2024, Intel stated that each of its photomasks holds ∼5 Petabyte of data and that it expects to increase this figure by a factor of 10x in the coming years. Partly this is due to the emergence of increasingly complex curvilinear mask pattern geometries associated with EUV, which require a much larger number of polygon vertices for accurate representation.(Shin et al., 2023) Data from Samsung shows a ¿4x increase in average mask data volume between the beginning of the EUV era around 2015 and 2023. Similar data from experts at IMS Nanofabrication and Nippon Control Systems suggests a 5x increase in file size since the EUV era and anticipates further increases due to application of EUV ILT.(Zillner et al., 2021)

The TAT for fracturing time depends on mask complexity, the fracture algorithm, and the hardware infrastructure employed(Shin et al., 2023), so, it is difficult to compare concrete time estimates.[49] Survey data from industry shows that average mask data preparation times have increased from 3.44 hours for nodes ≥130nm to 17.26 hours at nodes ≥7nm and

[41] Gilgenkrantz et. al. (2024)

[42] https://en.wikipedia.org/wiki/Proximity$_{e}ffect_{(}electron_{b}eam_{l}ithography)$

[43] https://nanolithography.gatech.edu/proximity.pdf

[44] The e-Beam Initiative uses the following definition: “MPC is defined as offline manipulation of geometry and/or dose of mask shapes during mask data preparation of the specified mask shapes received from OPC/ILT in order to more reliably manufacture the specified mask shapes on the physical mask or to maintain site-to-site compatibility.”

[45] Fujimura also states that “Mask process correction is the mask version of OPC or ILT.”

[46] For more background see Bork and Buck (2017), Bork and Buck (2019), Bork et. al. (2018), Bork et. al. (2021), and this video with Bork at Calibre.

[47] Similar corrections as MPC were already done at older nodes (PEC, LEC, and FEC). These targeted effects at longer length scales (“long-range”), while MPC targets effects at shorter length scales (“short-range”). These long-range corrections were integrated in the e-Beam writing process. (Bork and Buck (2019))

[48] https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11908/1190804/Model-based-mask-process-correction-for-EUV-Mask/10.1117/12.2601855.full

[49] For example, Bajpai (2021) give combined MPC/MDP runtimes of between a few minutes and 18 hours. Zillner (2021) gives a TAT of 5 hours using the OASIS.MBW 2.1 mask format, but it is unclear at which node size and CPU investment.

¡11nm.[5051] An expert from GlobalFoundries estimates that fracturing for a 12nm chip took only 1–2 hours[52]

A.2.2. PHOTOMASK MANUFACTURING

Following the transformation of the chip design data into an adequate file format, the physical manufacturing of the photomask can start. The transformed data feeds directly into the lithographic tools that are used to write the chip patterns on a photomask blank. Once the patterns are written, they are inscribed into the photomask through photoresist development and etching, followed by cleaning and application of a protective membrane.

Physical manufacturing of photomasks can be a significant contributor to photomask production TAT. As design complexity increases, the throughput of many types of mask writers and inspection tools has increased beyond an acceptable time limit. However, recent technological advancements have addressed some of these limitations, leading to permanently faster turnaround times, even for future technology nodes.

The most time-consuming activities of physical photomask manufacturing can be roughly divided into two categories:

- Mask writing and pattern transfer: writing the chip pattern onto the photomask blank surface and inscribing the pattern though photoresist development and etching
- Mask inspection and metrology: ensuring that photomasks adhere to stringent quality requirements, by detecting and eliminating defects and deviations from quality standards

Mask writing and pattern transfer[53]

A mask writer (or pattern generator) writes a chip pattern on top of a mask blank[5455] that is coated with photoresist.[56] EUV blanks consist of 40–50 alternating layers of silicon and molybdenum, capped by a ruthenium-based layer and a photoresist based on tantalum. The photoresist is developed,[57] selectively uncovering the underlying blank in the form of a chip pattern.[58] This pattern is then etched into the surface of the blank.[59] The remaining photoresist is then removed.[60] Two types of pattern generators are in use: laser and electron beam. Laser beam patterning is used for later features and non-critical layers,[61] while e-beam patterning is used for smaller features and critical layers.[62] Within e-beam pattern generators, the most common type is variable shape beam (VSB), but recently, the increased mask complexity caused by ILT has led to a shift toward multi-beam writers.[6364]

Mask writing is historically cited as one of the most time-consuming aspects of photomask production.[6566] In 2012, the e-Beam survey respondents cited “Write Time Reduction” as a core research priority.[67] In 2017, a researcher by Dai Nippon Printing cited write times as one of the reasons for increased mask turnaround time. This is because VSB mask writers become progressively slower as pattern complexity increases, since they expose the mask sequentially using a single moving beam that writes one shape at a time. However, the arrival of multi-beam mask writers has changed the picture and write times seem less problematic now. Multi-beam mask writers are faster than VSB writers because they use multiple electron

[50] https://www.ebeam.org/docs/pmj2021-ebeam-survey-results.pdf
[51] Sample size between 7 and 3, with the starting point defined as RET output.
[52] Expert 5
[53] https://photosciences.com/photomask-writing-and-development/
[54] Blank production could also be included here, but has historically not been a bottleneck
[55] https://www.halbleiter.org/en/photolithography/photomasks/
[56] https://photosciences.com/photomask-writing-and-development/
[57] https://photosciences.com/photomask-writing-and-development/
[58] https://www.halbleiter.org/en/photolithography/photomasks/
[59] https://photosciences.com/photomask-writing-and-development/
[60] https://semiengineering.com/euv-pellicle-uptime-and-resist-issues-continue/
[61] https://semiengineering.com/mask-lithography-issues-for-mature-nodes/
[62] https://semiengineering.com/next-gen-mask-writer-race-begins/
[63] https://semiengineering.com/next-gen-mask-writer-race-begins/
[64] Although there has also been research on improving VSB write times.
[65] https://semiengineering.com/battling-fab-cycle-times/
[66] Development, etching and photoresist removal are not frequently cited as major bottlenecks, although etching of EUV poses some unique challenges: blanks are more difficult to etch and new photoresists need to be developed.
[67] https://www.ebeam.org/docs/eBeamInitiative.2012survey.web.final.pdf

beams in parallel to expose the blank using a rasterized bitmap, rather than a single beam that exposes sequentially.[68] Leo Pang of D2S claims that "with the introduction of multi-beam mask writers, one of the major obstacles to full-chip ILT – excessive mask write time – was removed" (Pang et al., 2021).[69][70] Dai Nippon Printing, a large merchant photomask shop, claims that multi-beam writers reduce write time from 18+ hours to approximately 10 hours for cutting-edge masks.[71] The 2019/2020 eBeam Initiative survey claims that the average multi-beam write time was 12.14 hours with only 0.2% of masks being produced that way[72] Average VSB write time is at 7.91h and Laser write time at 2.33h.[73] The more recently introduced MBM-3000 multi-beam mask writer (MBM-3000) achieves full-mask writing times of 10–11 hours.(Matsumoto et al., 2024) [74] Expert estimates vary depending on chip type and technology node. Two experts from GlobalFoundries and TSMC agree with the eBeam Initiative survey that mask writing takes ∼10–12 hours.[75] Estimates for production time of an entire mask set are ∼3–4 weeks for a 5nm Intel GPU, ∼3 weeks for a Meta 5nm server chip, and ∼7–8 weeks for a 12nm chip at GlobalFoundries.[76]

Mask inspection, metrology, and repair

Photomasks need to adhere to extremely stringent quality requirements, as even small defects and imperfections can have a significant impact on the lithography process and the final chip performance. To ensure that photomasks meet these requirements, extensive measurements are performed throughout the manufacturing cycle. If defects are detected, they are either repaired in a way that maintains the mask's quality, or else, rejected. Following successful repair, the photomasks are cleaned and covered with a thin membrane, called pellicle, to avoid further particle contamination.[77][78] Mask inspection aims to detect defects on the photomask surface, such as contamination by dust particles (soft defects), errors in the pattern itself (hard defects),[79] deviations from the critical dimension (CD), as well as imperfections in mask flatness, image placement (registration), and mask-to-mask overlay.(Maniyara et al., 2017) Multiple[80] inspection steps take place at the mask shop: two before the pellicle (pre-pellicle) is attached and one after (post-pellicle). According to Sugimori et al. (2021) the initial inspection step, following mask writing, primarily targets hard defects. (Sugimori et al., 2021) After repairing these defects, a second inspection step is conducted to verify the repair's effectiveness and identify any soft defects. This second step also involves measuring deviations from key dimensional parameters, including line width, critical dimension uniformity (CDU), pitch, height, sidewall angle, and line edge/width roughness. The last inspection step takes place after pellicle attachment and focuses again on soft defects.[82] Three types of inspection tools are used for inspection of hard and soft defects: optical inspection tools,[83] electron beam scanners, and actinic inspection tools. For the most cutting-edge EUV masks, optical tools are not sufficient, as their resolution only allows for inspection at 7nm and higher.(Sugimori et al., 2021) E-beam inspection

[68] https://semiengineering.com/inspecting-patterning-euv-masks/

[69] While it would still be possible to use VSB at 7nm, according to IMS Nanofabrication this can result in operationally impractical pattering times up to 60 hours. Fujimura et.al. (2010) claim that mask write times exceeding 40 hours are infeasible for manufacturing, and that write times longer that 8-12 hours are operationally difficult for mask shops that handle a wide variety of masks. This may also be due to budgetary reasons. However, long write times may be more acceptable for high-volume chips. Platzgummer (2018) claims that writing times longer than two days may run into aging effects. Nakayamada (2013) mentions that a mask writing runtime of 10h - 1 day would be acceptable in 2023 (from 2013 perspective).

[70] Nevertheless, some complications remain. An expert claimed in 2022 that write times may continue to increase in the future, as the mask writer's data path could become saturated.

[71] https://www.global.dnp/news/detail/20167043_4126.html

[72] https://www.ebeam.org/docs/eBeam-Mask-Maker-Survey-2020.pdf

[73] IMS claimed in 2018 that multi-beam mask writers also have a 2-3x throughput advantage for more mature nodes.

[74] When interpreting these time estimates, it is also important to consider whether double patterning is used or not.

[75] Expert 5, Expert 4

[76] Expert 1, Expert 3, Expert 5

[77] https://semiengineering.com/searching-for-euv-mask-defects/

[78] As of 2017, cleaning was done with sulfuric acid and hydrogen peroxide, as well as alternative, non-wet cleaning systems. ASML's EUV pellicle consisted of polysilicon and was 50nm thick.

[79] https://semiengineering.com/photomask-shortages-grow-at-mature-nodes/

[80] Moreover, prior to the mask shop,[81] the mask blanks are inspected by the mask blank supplier.

[82] https://www.euvlitho.com/2020/S2.pdf

[83] https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11855/118550C/Study-of-high-throughput-EUV-mask-pattern-inspection-technologies-using/10.1117/12.2600987.short#_=_

has sufficient resolution for inspection down to the 1nm range,[84] but is much slower than optical inspection. [85][86] Actinic inspection (i.e. inspection that uses light of the same wavelength as in the lithography step) was introduced most recently, and has the advantage of sufficient defect sensitivity for EUV, high throughput, as well as the ability to inspect post-pellicle without the need for temporary pellicle removal. An example for an actinic inspection tool is Zeiss AIMS EUV. (Sugimori et al., 2021)[87][88] There are a number[89] of tools for inspection of dimensional parameters (metrology). Measurements of the critical dimensions are most often conducted with scanning electron microscopy (SEM), but increasing complexity of device geometries has led to other approaches, such as atomic force microscopy (AFM), and hybrid approaches.[90] There are also dedicated tools for ensuring correct pattern alignment on the photomask relative to predefined markers, and relative to patterns from previous lithography steps (overlay).[91] Examples for such tools are Zeiss PROVE, Zeiss ForTune, and KLA LMS IPRO.[92] Some tools combine multiple of these functions.[93] If inspection detects a defect, the mask shop tries to repair the defects as much as possible, as discarding masks is expensive. According to Lapedus (2022), there are two types of repair tools for 3nm and beyond: e-beam and nanomachining.[94] E-beam repair tools, like Zeiss MeRit,[95] eliminate the mask defect by targeting it with an electron beam. Nanomachining tools, like the Bruker nm-VI[96], eliminate defects using an AFM-guided diamond tip.[97] Both the Zeiss and Bruker tools also have particle removal capabilities. Overall, the number of defects on photomasks has increased with smaller node sizes. (Schneider et al., 2020)(Egodage et al., 2019)The increased use of RET has also increased the number of lithographically insignificant defects, which made it more difficult to detect lithographically significant defects.[98] Still, in 2020, only 35% of masks needed repair, and yield for EUV masks was at 91%. Only 0.19% of masks returned to the mask shop from the fab, with soft defects, hard defects, mask data prep errors, and OPC/ILT errors causing $\sim$75% of returns.[99]

It is difficult to find recent estimates for the total turnaround time for inspection, metrology, and repair. Mark Lapedus of Semiconductor Engineering claimed in 2016 that inspection and metrology combined can take up to 18 hours, roughly twice as much as "several years" before that.[100] In 2010, two experts from Semiconductor Digest claim that scan times for mask inspection are less than 6 hours, citing a TAT of 2 hours for optical, 3 hours for e-beam, and 30–80 hours [101] for e-beam (all 22nm). They expected scan times for optical and actinic to stay roughly constant, but expect e-beam TAT to increase by a factor of 4 with each successive node.[102] In 2017, Lapedus claimed that it is the "hope" that multi-beam inspection tools will be able to inspect at a xTAT of 3–5 hours – according to him, the same speed as optical inspection tools.[103] In 2021, Sugimori et al. claimed that they will release a multi-beam tool with a TAT of 6 hours per mask in 2022. NuFlare claims that its mask inspection system NPI-8000 for 10nm/7nm has an inspection time of 1 hour, but does not specify what is included in this time. [104](Sugimori et al., 2021)

[84] https://semiengineering.com/knowledge$_{c}enters/manufacturing/process/wafer-inspection/e-beam-inspection/$

[85] Although there are efforts to develop high-throughput , multi-e-beam inspection tools. (Sugimori et. al. (2021))

[86] https://semiengineering.com/knowledge$_{c}enters/manufacturing/process/wafer-inspection/e-beam-inspection/$

[87] https://www.zeiss.com/semiconductor-manufacturing-technology/products/photomask-solutions/mask-qualification.html

[88] Optical inspection is not possible here, as the pellicle is opaque to 193nm light. Actinic inspection is possible with the pellicle still attached.

[89] https://semiengineering.com/knowledge$_{c}enters/manufacturing/process/metrology/$

[90] https://semiengineering.com/knowledge$_{c}enters/manufacturing/process/metrology/cd-sem/$

[91] https://www.photomaskportal.com/terminology.html

[92] https://www.zeiss.de/semiconductor-manufacturing-technology/produkte/photomaskenloesungen/maskenmetrologie.html

[93] https://www.researchgate.net/publication/252779712$_{P}ROVE_{t}he_{n}ext_{g}eneration_{r}egistration_{m}etrology_{t}ool_{s}tatus_{r}eport$

[94] https://semiengineering.com/photomask-shortages-grow-at-mature-nodes/

[95] https://www.zeiss.de/semiconductor-manufacturing-technology/produkte/photomaskenloesungen/maskenreparatur.html

[96] https://www.bruker.com/de/products-and-solutions/semiconductor-solutions/photomask-repair/nm-vi.html

[97] https://semiengineering.com/photomask-shortages-grow-at-mature-nodes/

[98] https://semiengineering.com/challenges-mount-for-photomasks/

[99] https://www.ebeam.org/docs/eBeam-Mask-Maker-Survey-2020.pdf

[100] https://semiengineering.com/taming-mask-metrology/

[101] They expect this to go up to 320-640 hours in 2020, although it is unclear if this prediction came true. Semiconductor Engineering claims that e-beam inspection of a full wafer can take up to several days, although it is unclear how comparable this is to photomasks.

[102] https://sst.semiconductor-digest.com/2010/08/mask-and$_{t}emplate/$

[103] https://semiengineering.com/searching-for-euv-mask-defects/

[104] https://www.nuflare.co.jp/english/products/mask/

A.3. Prototyping

The finished photomasks are delivered to the fab, where a small batch of chip prototypes (or engineering samples) is produced.[105] Once available, these samples undergo extensive testing and validation to make sure that the chip logic functions as intended by the designers, and that the non-logic parts of the chip perform adequately under a wide range of conditions (post-silicon validation). If defects are found, the validation teams try to identify the root cause behind these defects and try to eliminate them via debugging. Often this is not possible, so that the chip has to go back to the design phase (respin) and another prototype is manufactured.

Prototyping involves very time-intensive steps. It is well-known that cutting-edge chips take several months to manufacture. For example, a 5nm TSMC chip with 81 layers[106] and a cycle time of 1–1.2 days per mask layer[107] would take at least 80–96 days to manufacture [108]. Hence, for this chip, the design firm would have to wait approximately 3 months until they have a prototype ready for testing. Inspection times are somewhat harder to gauge, but anecdotal evidence from industry suggests that it often takes months to complete.

In line with the description above, I will group the most time-consuming prototyping activities into three clusters:

- Engineering sample manufacturing: producing an initial set of prototype wafers to enable post-silicon validation of the chip design and to test the manufacturing process
- Post-silicon validation: inspecting the engineering sample to detect and eliminate defects prior to mass production
- Respins: discarding the existing engineering sample and iterating the chip design in response to irreparable defects

Engineering sample manufacturing

When the mask set is completed, a set of prototype wafers is produced (engineering samples). This initial set of wafers (first silicon) is used to validate that a chip design is compliant with specifications, both in terms of performance and manufacturability. If bugs are found in the prototype wafers, then this gives chip designers the chance to go back to the drawing board for design modifications, and fabs the opportunity to modify the production processes.[109] For prototype manufacturing within an existing node, it is common for chip design firms to use multi-product wafer services.[110] [111]These are manufacturing arrangements within fabs that allow for multiple chip designs by different clients to be printed on a single wafer.[112] This makes it cheaper for individual design firms to manufacture prototypes, as costs for manufacturing and photomasks are shared among several clients. All three of the major chip manufacturing firms TSMC, Samsung, and Intel have multi-product wafer services.[113][114][115] When a new product enters prototype production, the fab ensures that its production lines are adequately prepared. It starts by sending small number[116] of lots through the production line (pilot lots or pipe-cleaner lots), which are used to calibrate the manufacturing tools and processes.[117] [118]Moreover, fabs may run experimental lots meant to simulate variations in process parameters that may occur during production (corner lots). This involves splitting a set of wafers into different lots, and varying the manufacturing parameters between these lots.[119] Once the line is prepared, the wafers go through a large number of manufacturing steps, such as deposition, lithography, etching, cleaning, polishing, ion implantation, and metrology. These processes are well-documented, and a more detailed

[105]Depending on the chip design and the preferences of the design firm, it is not always necessary to have the full photomask set ready before wafer manufacturing can start. Manufacturers often choose to start manufacturing already when the first few photomasks arrive at the fab, in order to save time.

[106]https://semiengineering.com/whats-next-for-euv/

[107]https://esg.tsmc.com/download/file/2018$_{t}smc_{c}sr_{r}eport_{p}ublished_{M}ay_{2}019/english/pdf/e_{4}innovationAndService.pdf$

[108]This is just a BOTEC based on likely outdated numbers for both mask count and days per mask layer.

[109]Expert 8

[110]https://en.wikipedia.org/wiki/Multi-project$_{w}afer_{s}ervice$

[111]Expert 9

[112]https://en.wikipedia.org/wiki/Multi-project$_{w}afer_{s}ervice$

[113]https://www.intel.com/content/www/us/en/foundry/manufacturing/shuttle.html

[114]https://semiconductor.samsung.com/foundry/manufacturing/mpw-service/

[115]https://www.tsmc.com/english/dedicatedFoundry/services/cyberShuttle

[116]Expert 8 mentioned 3 lots in his example.

[117]Expert 8

[118]Expert 8 mentioned that production may already start if not all photomasks are ready, to save time.

[119]Expert 9

description is out of scope here. The output of this process is a set of wafers that the chip design firm uses for validation and design iteration. Multiple lots of engineering samples pass through the fab simultaneously, with staggered start times. This is helpful for shortening iteration times: if a defect is detected in one of the earlier engineering samples, then the fab can adjust the production parameters mid-process for later samples, so that it is not necessary to re-run a wafer from scratch.[120] There is a set of feedback loops along the design-to-manufacturing lifecycle. This is also known as design-technology co-optimization (DTCO). Effective DTCO requires efficient information exchange between the different steps along the life cycle.[121]

The turnaround time for engineering samples depends on the cycle time of the fab, which in turn depends on the number of mask layers and the number of days a fab takes to complete a mask layer (days per mask layer, or DPML). The number of mask layers depends on the technology node and the complexity of the product. For example, TSMC is using 87 mask layers for the 7nm DUV process, 79 mask layers for the 7nm EUV process, of which 4 are EUV, and 81 mask layers for the 5nm process, 14 of which are using EUV.[122] The days per mask layer depends on the priority level desired by the client. Fabs offer hot lot runs that reduce cycle time by up to 40%, provided that the client is willing to pay for this. A 2009 SEC filing for TSMC suggests that regular DPML is between 1.0 and 1.4 DPML, while hot lots would be processed at 0.8 DPML.[123] An expert from Intel claims that for Intel Arc, DPML was around 0.8–0.9 DPML, with up to 0.5 DPML for the fastest hot lot runs. An expert from Intel estimates that "traditional GPU or complex SoC" have around ∼100–120 mask layers, which Intel is producing at a speed of ∼0.6–0.9 DPML, which would imply a cycle time of ∼60–108 days.[124] By contrast, the 5nm Intel Xeon with ∼50 mask layers was produced at ∼1.5–2 DPML, implying a cycle time of ∼75–100 days.[125] At Meta, a 3nm chip would take ∼100–120 days from first wafer start to first silicon, although 100 days would constitute an expensive acceleration.[126] Finally, for a 12nm chip at GlobalFoundries, completion of the engineering sample took ∼45–50 days, although this could be expedited to ∼30–35 days.[127]

Post-silicon validation

Post-silicon validation (PSV) starts once an engineering sample of the final chip is available. The purpose of post-silicon validation is to ensure that the chip is free of defects and ready for mass production. It involves tests that ensure that there are no flaws in the chip's logic design, and that the chip is ready to be deployed in a real-world environment.[128] These tests are conducted directly on the silicon. Ideally, a chip passess all tests on the first pass and can directly go into mass production[129] without major bug-fixing and workarounds. If major defects are discovered, then this might require a costly design iteration, or respin, during which the chip may be re-designed, re-manufactured, and re-validated. Mishra (2017) mentions five different types of post-silicon validation:[130] (Mishra et al., 2017)

- Power-on debug: switching on the device for the first time using a customisable debug board. If the device does not power on at the first try, it may be necessary to simplify the system configuration, and to add more complexity over time until a stable power-on setup is found.
- Basic hardware logic validation: verifying that the logic circuits in the chip work as desired. This involves different types of software tests, supported by specialized hardware that makes the internals of the chip more observable and controllable.
- Hardware/software compatibility validation: checking how well the chip works with external hardware, different operating systems and applications, as well as network protocols and communication infrastructure.

[120] Expert 8

[121] https://semiwiki.com/events/339386-spie-2023-buzz-siemens-aims-to-break-down-innovation-barriers-by-extending-design-technology-co-optimization/

[122] https://www.3dincites.com/2021/08/sustainability-in-the-semiconductor-fab-and-sub-fab/

[123] https://www.sec.gov/Archives/edgar/data/703361/000070336112000032/exhibit1036idt-tsmcfoundry.htm: :text=(a)%20TSMC%20shall%20use%20comm

[124] Expert 1

[125] Expert 2

[126] Expert 3

[127] Expert 5

[128] https://www.tessolve.com/post-silicon-validation/

[129] The decision to release a chip to mass production is also called Product Release Qualification (PRQ). The timeline for this decision depends on the results of PSV, but is also influenced by economic incentives to launch a product as quickly as feasible. (Mishra (2017))

[130] This list is not exhaustive, and does not specifically refer to GPU development. An Apple job ad for a GPU post-silicon validation engineer seems to match at least some of these activities.

- Electrical validation: testing the electrical characteristics of the chip, for example I/O, power delivery, clock, and analog/mixed-signal components. Goal is to ensure that it performs well, even under extreme operating conditions.
- Speed-path validation: identifying bottlenecks in the circuit design, such as slow transistors or limitations in the execution cycle.

Ziv (2019) mentions a general framework for execution of various types of validation: (Ziv, Accessed on 2025-05-12) [131] [132][133]

- Stimuli generation: generating stimuli that exercise various components within the chip
- Checking: checking that the chip behaves as expected and detecting when it does not
- Coverage: ensuring that all aspects and features of the chip are verified
- Debug: identifying and fixing the root causes of defects

Compared to the pre-silicon validation that happens before tapeout, PSV has the advantage that tests can be run at the actual clock speed of the processor, so that functional tests can be executed much more rapidly. In pre-silicon validation, tests rely on detailed RTL simulations that run about a billion times slower than the real system. Moreover, PSV allows designers to test non-functional characteristics (e.g. electricity) of the chip, which is not possible with a simulation. A downside of PSV is that – unlike in a simulation – there is no complete visibility into the internals of a silicon chip. Any signals to be observed must be routed through a measuring device, so internal visibility is limited by the number of such devices.[134] Moreover, internal signals cannot be controlled perfectly, as the chip architecture only allows for a limited number of configuration options. Both factors make it more difficult to identify and isolate defects. Both pre- and post-silicon validation are necessary and complement each other. There has been a general trend to analyze characteristics of a chip as early as possible in the development flow (shift left), leading to extensive pre-silicon simulations,[135] but post-silicon validation remains important. Several experts[136] (Mishra et al., 2017) also emphasize the connection between the two validation steps, for example when post-silicon test cases are generated using pre-silicon environments.

Validation time makes up a significant share of overall project time. In 2022, more than 50% of surveyed IC/ASIC design projects claimed that ¿50% of their project time was spent on validation. Across all projects surveyed, the mean peak-time engineering headcount ratio between design and validation was 1:1.[137] [138] Moreover, a number of experts(Mishra et al., 2017) (Nahir et al., 2010) claim that post-silicon validation makes ¿50% of overall design cost. The exact amount of time spent on validation time is somewhat unpredictable[139] and varies between different projects. For example, less time is typically spent on validation if the design makes use of existing, pre-verified IP.[140] It is difficult to find specific data on debug time for cutting-edge AI chips, or GPU more generally. A few bits of anecdotal data suggest that validation can take weeks to months. Mishra et al. (2017) claim that achieving a stable power-on debug alone can take several weeks[141] and that they spent at least 3 months debugging an Intel POWER8 CPU, although it is unclear how comparable this is to GPU-debugging in industry settings.(Mishra et al., 2017) [142] A professor from Carnegie Mellon University called post-silicon debug a "dirty little secret" that can cost $15 million to $20 million and take ∼6 months to complete, although this information is

[131] https://theory.stanford.edu/ barrett/fmcad/slides/3$_{Ziv.pdf}$

[132] This framework applies both pre-silicon and post-silicon, but it is unclear if it is applicable to all types of validation (e.g. electrical, power-on) or just logic validation.

[133] Related frameworks are provided by Mitra et. al. (2010) and Semiengineering.

[134] Chips are already designed to facilitate post-silicon debug and validation (also called Design-for-debug or DfD). A key focus is improving observability of internal states of the system, through on-chip instrumentation like scan chains and signal tracers. In some cases, the DfD hardware features may take up more than 20% of silicon real estate. (Mishra (2017)).

[135] https://semiengineering.com/knowledge$_{centers/eda-design/methodologies-and-flows/shift-left/}$

[136] https://semiengineering.com/transforming-silicon-bring-up/

[137] The most time-consuming tasks were test planning, test creation/simulation, and testbench development.

[138] Expert 3 mentioned that for an Intel Xeon 5nm, 200-300 people worked on post-silicon validation.

[139] https://semiengineering.com/reduction-in-first-silicon-success/

[140] 2022 Wilson Research Group IC/ASIC functional verification trends

[141] Intel claims that they have shortened this time to just a few hours since then.

[142] They also list 300 debugging days on one of their graphs, but it is unclear if this is the actual turnaround time.

from 2007 and may be outdated.[143] An article on post-silicon debugging from 2010 mentions that it took ∼3 months of randomized simulations to uncover one particular defect.[144] Experts at Intel estimate[145] that post-silicon validation of a chip takes between ∼3–6 months, with GPUs usually one the longer side due to memory-access-intensive tensor cores.[146] At Meta, validation of an unspecified cutting-edge chip was reported to take ∼20 weeks.[147] Even longer cycle times were reported at GlobalFoundries, where an expert claims that validation of an (unspecified) chip can took ∼9 months.[148]

Respins

If defects are detected in a chip prototype, and these defects are not easily addressable, the chip design firm may decide to go back to the drawing board to iterate the chip design.[149] This means discarding the current silicon prototype, modifying the design file to remove defects, writing new photomasks, manufacturing a new prototype, and checking again for defects. This procedure is repeated until the chip has reached an acceptable level of quality and is ready for product release qualification. Firms try to keep the number of respins as low as possible. Printing new masks and re-running many of the engineering and manufacturing steps is expensive, and causes a significant delay in time-to-market. Nevertheless, first silicon success is not the norm. A study by the Wilson Research Group shows that in 2022, only ∼25% of IC/ASIC needed no respins, while ∼45% needed 1 respin, ∼20% needed 2 respins, while ∼10% needed 3 or more. The most common flaws leading to respins were logic defects, followed by analog, power consumption, yield, clocking, mixed-signal interface, and crosstalk. The most frequent root causes of logic defects were design errors, changes in specification and incorrect/incomplete specification.[150] Between 2016 and 2022 the percentage of projects that succeeded at first silicon decreased steadily, from ∼33% to ∼23%. Bailey (2024) attributes this development to a mixture of new technological issues, such as thermal requirements and security features, and a shortage in qualified validation engineers. [151]

The additional time investment for a respin is not necessarily equal to the time it takes to manufacture a new prototype. As mentioned earlier, engineering samples pass through the fab with staggered starting times, so that potential defects can be eliminated by process adjustments mid-line. Hence, if a respin is initiated,[152] the validation team does not necessarily need to wait until a new engineering sample is finished, but can adjust one of the half-finished samples that are on the way. For example, if chip manufacturing time is 3 months, then the time to get a new engineering sample may only be 1.5 months.[153] An expert from Intel mentioned that the median time investment for a respin is 4–6 weeks, with the most complex ones taking 10–12 weeks.[154] An expert at Meta gave a somewhat higher estimate, at ∼3–5 months.

A.4. High-volume manufacturing

After a new chip meets all specifications necessary for product release qualification, it is ready to be transferred to high-volume manufacturing (HVM). At this stage, less testing is required, as process parameters are optimized, and fab engineers now need to ensure that wafer production does not deviate too much from these parameters. There may be some yield improvements between the engineering samples and the fully matured HVM wafers. According to an expert[155], engineering sample lots tend to have yields between 90–95% while HVM lots tend to yield between 95–97%, but it is unclear how representative this is across nodes and products.

This yield improvement implies that there is some time delay until HVM is running at full efficiency after product release qualification. Expert estimates for ramp to HVM vary between firms and product types. One expert who worked on Intel

[143] https://www.eetimes.com/post-silicon-debugging-worth-a-second-look//

[144] https://www.elektronikpraxis.de/lueckenlose-pruefung-komplexer-chip-designs-im-post-silicon-debugging-durch-tools-der-formalen-verifikation-a-250509/?p=3

[145] Expert 1

[146] Expert 1 Expert 2

[147] Expert 3

[148] Expert 5

[149] Respins may also happen due to last-minute changes in the chip specification that are unrelated to defects, for example, to stay competitive with rival products that enter the market at the same time.

[150] https://resources.sw.siemens.com/en-US/white-paper-2022-wilson-research-group-functional-verification-study-ic-asic-functional-verification-trend-report

[151] https://semiengineering.com/trouble-ahead-for-ic-verification/

[152] Expert 8 has stated that the decision to initiate a respin may be taken 1-2 months into validation.

[153] Expert 8

[154] Expert 1

[155] Expert 11

5nm GPU claims that there are only between ∼1.5–2 months between completion of PSV and start of HVM. Another Intel expert contradicts, claiming that HVM can only start ∼6–12 months after PSV, due to additional system-level testing and client validation steps. Other experts at Meta, TSMC, and GlobalFoundries give time estimates between several weeks and 3 months. It is not entirely clear how to reconcile these differences. Possibly, this is the result of differences in client requirements both between products in the same firm, and between custom and non-custom silicon manufacturers.

In some cases, fabs start this process while engineering samples are still undergoing validation – this is called risk production.[156] This allows design firms to ship to customers faster, but carries the risk that shipped chips contain defects. For example, TSMC is expected to start risk production for its 2nm node in Q4/2024.[157]

It is possible that HVM fails despite a chip passing product release qualification successfully. For example, a chip in Intel's 10nm process passed tape-out successfully, but ran into significant delays during its HMV ramp[158] due to a problem with its cobalt interconnect. The defect was too subtle to be detected at an engineering sample size of 100 wafers, but became apparent at higher volumes. This defect delayed the product launch by 2 years.

B. Time estimates of process steps

B.1. Overview

Time estimates for production steps between tapeout and high-volume manufacturing vary depending on a number of factors, such as node, design complexity, number of (EUV) layers, market requirements, and company practices. To get a balanced view that does not hinge too much on the specifics of one particular product or company, I interviewed 5 experts at 4 different companies to get specific time estimates from them.

For this purpose, I divide the tapeout-to-HVM process into 5 sequential steps:

1. Photomask production
2. Engineering sample
3. Post-silicon validation
4. Respins
5. Ramp to HVM

For each category, I asked 5 experts from Intel, Meta, TSMC, and GlobalFoundries to estimate how long this step would take, based on the most-cutting-edge, most GPU-like chip they worked with. Expert profiles are listed in the Appendix.

Our interviews showed that the overall range for getting an existing node into production varies between ∼7–15.8 months. The table below shows the time estimates in more detail.

When interpreting these results, the following caveats should be noted:

- The separation into sequential steps is an imperfect approximation. Steps may be overlapping. For example, wafer production may already start once the first few masks are complete, before the entire mask set is completed.
- Steps may be executed several times. For example, it is common that 1–2 respins are necessary, but some chips require no respins at all.
- Steps are not clearly and uniformly defined across companies. Companies use different words for similar steps, and if they use the same word, they may refer to different sets of processes. During the interviews, there was often no time to fully align on terminology.
- The experts I interviewed did not work on the most cutting-edge AI chips. It was difficult to find people with the necessary expertise who were willing to share information. However some of the experts I interviewed have experience

[156] According to Expert 1, the decision to initiate risk production may already be taken 1 month into post-silicon validation.
[157] https://www.digitimes.com/news/a20240219PD225/tsmc-2nm-production-2024.html: :text=TSMC
[158] Expert 6

in working with GPUs (e.g. Intel Arc) and with GPU-related products, such as Intel server chips (e.g. Intel Xeon). Moreover, all but one had experience working on either 5nm or 3nm.

- Experts have incomplete knowledge, and may be overconfident. The quality of their insights is probably higher in areas where they have particular expertise. I tried to mark in the table whenever I had a sense that experts were overconfident, or seemed uncertain.

B.2. Summary of time estimates

Table 2: Summary of time estimates for NPI process steps.

Step	Time estimate	Comments	Source
Photomask production	1–1.5 months	Of which 1–2 weeks for OPC-type processes, and 3–4 weeks for mask writing.	Expert 1 (Intel)
	0.5–0.75 months[159]	Photomask manufacturing is pipelined, but happens largely in parallel. Production can start once the first couple of masks arrive at the fab. For example, if all photomasks for the transistors are complete, one can commence with base-layer tapeout, followed by metal layer tapeout several days later.	Expert 2 (Intel)
	∼9–17 days until wafer production starts	Post-tapeout flow took around 1 week for products in 5nm–3nm at Marvell, Broadcom, and Meta, and a few days for 5nm at Intel. A shortening of the process to 2–3 days is possible. Masks are manufactured in parallel. Production time of 7–10 days for all base layers, then 2 weeks for metal layers. For a 5nm data center chip at Intel, it took 7–10 days to complete the first 80 masks (transistors and metal contact layers), then ∼2 weeks to complete the next 40 layers (metal layers).	Expert 3 (Meta)
	0.5–1 months[160]	Masks are fabricated in parallel, but it is unclear how many at the same time. For 7nm Apple A12 and AMD Ryzen 3000 mask fabrication and inspection took upwards of 4 weeks. ILT takes several days to a week. OPC takes 1–3 days, but can be done faster. MPC takes 1–2 days. Mask writing takes less than 24 hours, now down to 10–12 hours due to multi-beam writers. Mask fabrication and inspection jointly take less than one week.	Expert 4 (TSMC)

Continued on next page

[159] Expert 2

[160] Expert states that this is overlapping with the pre-tapeout process, unclear which steps are involved here.

Table 2: Summary of time estimates for NPI process steps (continued).

Step	Time estimate	Comments	Source
	∼2 days until wafer production starts	This was the goal at GlobalFoundries for established nodes up to, and including, 12nm. Photomasks were delivered at a schedule matching maximum lot speed. For example, it took 1 day from tapeout until photomask production could start. Then the first 4 masks were delivered within a day, followed by 1–2 masks per day until the first engineering lot was completed. Time for data preparation for the first few masks is 10 hours. Fracturing takes 1–2 hours of computing time. Mask writing takes up to 12 hours, but can be done faster if inspection is reduced. At a 12nm chip with 60 layers, it took 50–60 days until all masks were finished, but this can be expedited to 30 days, or even just a couple of days, as mask shops have enough capacity.	Expert 5 (GF)
Engineering sample	3–4 months	Engineering sample turnaround time depends directly on the number of layers of a chip, as well as the fab's throughput, measured in days per mask layer. The expert did not mention the specific number of mask layers for Discrete Graphics 2, but gave a ballpark of ∼100–120 days per mask layer for "traditional GPU or complex SoC", with a throughput estimate of 0.6–0.9 days per mask layer.	Expert 1 (Intel)
	2–3 months	This is an estimate for an Intel Xeon with 50 layers at 1.5–2 days per mask layer.[161] It also includes 2–3 weeks of dicing and packaging.	Expert 2 (Intel)
	4–5.3 months (3nm chip at Meta)	A 3nm chip at Meta would take ∼100–120 days from first wafer start to first silicon, where 120 days is more common and 100 days would be an expensive acceleration. Wafer production is followed by additional steps which add between 3–9 weeks on top of this estimate. These steps include bumping, [162]manufacturing testing,[163] and packaging, which take ∼3 weeks in total. If CoWoS is involved, this would add an additional 5–6 weeks for packaging.	Expert 3 (Meta)

Continued on next page

[161] With 2 days per mask layer, and 3 weeks of dicing and packaging, a 50-layer chip would take 4 months in total. It is unclear why the expert did not extend their range estimate to 4 months.

[162] Adding the metal pillars that go from the pads on the die to the packaged substrate.

[163] A limited test program that includes pin checks, brief power supply tests, and brief SRAM tests. The goal is to ensure that the bring-up equipment is not damaged.

Table 2: Summary of time estimates for NPI process steps (continued).

Step	Time estimate	Comments	Source
	3–4 months (3nm)	For N7 Apple A12, AMD Ryzen 3000 (∼80 layers), engineering sample fabrication and post-silicon validation jointly take 3–6 months, but is uncertain about this.	Expert 4 (TSMC)
	∼1–1.5 months (12nm, 60 layers)	At 12nm it took 45–50 days until an engineering sample was out, but this can be expedited to 30–35 days.	Expert 5 (GF)
Post-silicon validation	4–6 months	Expects a variation of around plus-minus 1 month in either direction, depending on the type of chip. GPU are the most complex to verify due to a larger number of memory-access-intensive tensor cores. Testing duration also depends on whether previous generations of a product were tested already.	Expert 1 (Intel)
	3–6 months	Functional validation and non-functional validation taking place in parallel.	Expert 2 (Intel)
	5.25–6 months	Bring-up takes ∼1–4 weeks, depending on business practices. AMD typically takes 4 weeks for bring-up, whereas Meta has a 1-week bring-up target. Validation itself takes ∼20 weeks. Custom silicon providers may be more free to take risks here, as they are less bound by customer requirements.	Expert 3 (Meta)
	3–4 months (3nm)	Claims that for a N7 Apple A12 or AMD Ryzen 3000 post-silicon validation would take only 2 months or less, but sounded uncertain.	Expert 4 (TSMC)
	4–9 months	The time between the first prototype and being ready to ramp production can be as low as 4–6 months[164], but this seems like an aggressive estimate. A more realistic estimate is 9 months. It is somewhat unclear if this refers to an older node, or a cutting-edge AI chip. Cutting-edge products take longer to verify, as they are less well understood.	Expert 5 (GF)
Respins	1–3 months	The expert expects that the median is around 4–6 weeks, with the most complex ones taking 10–12 weeks. The delay time depends on the type of defect. An interconnect issue may be easier to fix, as it only requires adjustments in one of the higher metal layers, so a new engineering sample is closer in the pipeline. If the defect lies deeper in the layer stack, it takes longer for the new engineering sample to arrive, as more additional layers need to be added.	Expert 1 (Intel)

Continued on next page

[164]The expert expressed some uncertainty around this estimate.

Table 2: Summary of time estimates for NPI process steps (continued).

Step	Time estimate	Comments	Source
	0–3 months[165]	Respin is initiated some time within the 3–6 month validation time, for example 1 month after arrival of the first engineering sample. If a defect is high in the metal layer stack, then time until the next engineering sample arrives may be cut by 50%–66% compared to the total manufacturing time.	Expert 2 (Intel)
	3–5 months	For NVIDIA and AMD at "several steppings", the exact number is unclear. Custom silicon manufacturers try to avoid respins in general. AMD and NVIDIA are different: they go into respins more often, because they promised certain features to their customers and cannot change specifications on the fly. Decision to respin is typically made after 16 weeks, but could be made earlier depending on budget.	Expert 3 (Meta)
	–		Expert 4 (TSMC)
	–		Expert 5 (GF)
Ramp to HVM	1.5–2 months[166] after PSV	There is not much delay between product release qualification and readiness for HVM. The PRQ already implies that defect density is low enough for HVM to start. Especially if the chip is in a mature node, the defect density should already be at HVM level. This means that full HVM volume will be reached around 3–4 months after PRQ is completed, potentially earlier if risk production is started successfully during validation.	Expert 1 (Intel)
	6–12 months after PSV	Chips go through two additional validations after PSV: (1) "System level testing", where chips are tested in a data center, and (2) Validation by the lead customer. Each of these validation steps take 1–2 quarters, coming on top of the 4–6 months necessary for PSV. This is somewhat surprising, as other experts had not mentioned these steps. It would imply that driving a chip to HVM takes much longer than $\sim$1 year. Indeed, the expert mentioned that it took Intel $\sim$3 years to get the Intel Xeon into mass production. I'm not sure to which degree this generalizes to TSMC and Samsung.	Expert 2 (Intel)

Continued on next page

[165] The expert did not state this number explicitly, this is inferred from other information they gave.

[166] More precisely: 6-7.5 months after first silicon. The 1.5-2 months is inferred by subtracting the time estimate for PSV.

Table 2: Summary of time estimates for NPI process steps (continued).

Step	Time estimate	Comments	Source
	6–9 months after bring-up	This is for first-silicon success. A typical timeline at NVIDIA or AMD would be 9–15 months from bring-up, depending on the number of respins. It is somewhat unclear what the additional time is used for, but partly, it seems to be a "manufacturing test program" which aims at meeting customer expectations in terms of yield (e.g. 500 defects per million units) and quality (e.g. 1000 hours of HTOL[167] testing).	Expert 3 (Meta)
	Several weeks	This time is needed to ensure that the recipes, tools, and processes used to produce the chip are configured according to customer standards and requirements.	Expert 4 (TSMC)
	2–3 months	Exact time depends on market requirements and varies between companies. Includes burn-in qualification. Ramping could happen directly after first silicon or even earlier, but this is risky.	Expert 5 (GF)
Total time (assuming no respins)	9.5–13.5 months	Assuming that photomask production is efficiently staggered, and wafer production starts just a few days after tapeout, the range would be 8.5–12.0 months.	Expert 1 (Intel)
	8.5–15.75 months	Time between PSV and HVM was calculated by subtracting the PSV time from the time between bring-up and HVM. Assuming that photomask production is efficiently staggered, and wafer production starts just a few days after tapeout, the range would be 8.0–15.0 months.	Expert 2 (Intel)
	10.3–14.8 months	Assuming that photomask production is efficiently staggered, and wafer production starts just a few days after tapeout, the range would be 10.0–14.3 months.	Expert 3 (Meta)
	7–10.5 months	This assumes that "several weeks" amounts to 0.5–1.5 months. Assuming that photomask production is efficiently staggered, and wafer production starts just a few days after tapeout, the range would be 7.7–14.5 months.	Expert 4 (TSMC)
	7–13.5 months	–	Expert 5 (GF)

[167] High-temperature operating life