Project 3: Semiconductor Wafer Production Queuing Model

By: James Tourkistas

Abstract

As demand for semiconductors grows and production capacity becomes increasingly strained, this paper considers how a queueing model might be used to guide capacity growth discussions and estimate potential impacts of adding different forms of production capacity. The proposed model uses a poisson distribution to estimate daily orders and order probabilities from actual demand to estimate likely order type. It then uses daily machine capacity estimates to process orders while tracking machine idle time, missed orders, wait time, and revenue.

Section 1: Background

Semiconductors, a.k.a. integrated circuits (IC), make up essential components of modern technology. Everything from medical devices, computers, artificial intelligence, cellphones, automobiles, the internet of things, tablets, and even game consoles require them to function. Since 2016, as modernization and advancement persists, demand (as revenue) has grown 7.9% annually.[1] From 2016 to 2021, sector revenue is expected to grow from $297.7 billion to $434.5 billion.[2] Increasing demand for some products left dependent industries facing serious supply issues. Geopolitical concerns, hoarding, quantum effects on state of the art designs, extremely high costs of entry/retooling, natural disasters, constant variation in demand for different/ evolving design types, have been restricting supply and causing production disruptions.[3] As of 2020, 90% of the state of the art IC production occurred in Taiwan and South Korea.[4] Concerns over intellectual property loss, foreign dependence, and security, have lead to renewed US government efforts to encourage domestic production through subsidies.[5]

The industry faces numerous supply side challenges. Significant R & D, machinery, and fabrication plant (fab) construction costs create barriers to entry and necessitate careful risk analysis for investments in additional production capacity. Estimated chip design costs range

---

[1] Platzer at 13.

[2] Platzer at 13.

[3] Jeong.

[4] Platzer at 12.

[5] Platzer at 14-15; 18-20; Vatas.

from $69 million for 22 nm features to $540 million for state of the art (5 nm) features. New fab construction can range from $1.1 billion (22 nm) to $5.4 billion for state of the art.[6] A fab shell can take 2 years to build and another 12 to 18 months to ramp up to full production.[7] Advanced logic fabs take 5 to 10 years to reach break even points, contingent on plant capacity utilization.[8] By some extreme estimates, fab construction costs between $7 to $20 billion with a useful life of 5 to 6 years.[9]

Money can be made but mistakes are costly. From 2015 to 2019, top industry participants have seen $85 million to $10.9 billion average annual profits while others lost $17.9 to $481 million.[10] 2019 forecasts from the top 15 producers ranged from $7.9 billion to $69.8 billion in sales.[11] The average annual labor cost per employee is $166,400 in 2019.[12] In 2018, for example, an estimated 20% of sales was spent on capital expenditures, including retooling and new facility production to keep up with evolving needs.[13] A circuit's processing power is directly related to the number of transistors printed on a silicone chip. Although rapidly approaching quantum/ atomic limits (.5 nm), over time, heavy R & D investment (15 to 20% of sales) has allowed the number double every 18 months to 2 years rendering produced stock obsolete and causing significant risk to investments.[14] From 1971 to 2008, transistor counts in processors have gone from 2,300 to 2,000,000,000.[15] The current state of the art is a 5 nm feature size.

According to the Semiconductor Industry Association, three examples of semiconductor production, include:

1) Type 1: 3,000 employee and 35,000 wafer per month capacity advanced logic (300mm wafers, 5 nm features) fab producing IC's for mobile/advanced computing processors, building and equipment. Type 1 is estimated to require an initial capex of 20 billion and 1 billion each year in operating expenses.

2) Type 2: A 6,000 employee and 100,000 wafer per month capacity fab producing IC's for advanced flash memory (300mm wafers, 20 nm features). Type 2 is

---

[6] Bauer at 5.

[7] Bauer at 5.

[8] Bauer at 6.

[9] Platzer at 14.

[10] Bauer; Platzer at 19.

[11] Platzer at 50.

[12] Platzer at 20; Barbe.

[13] Platzer at 14.

[14] Platzer at 4; VerWey at 16.

[15] VerWey at 16.

estimated to require an initial capex of 20 billion and 1.8 billion each year in operating expenses., and

3) Type 3: A 3,000 employee and 40,000 wafer per month capacity fab producing advanced analog (300mm wafers, 65 nm features) fabs electronics for vehicle electronics/renewable energy. Type 3 is estimated to require an initial capex of 5 billion and .8 billion each year in operating expenses.[16]

Industry participants have asked the US government to increase available tax incentives from 10 to 15% to 25-40% in line with asian countries.[17] These incentives can amount to .1 to 1 billion in annual savings. [18] This paper proposes the use of an adapted queuing model as a mechanism to estimate the effects of increasing various types of production capacity.

Section 2: Model Description

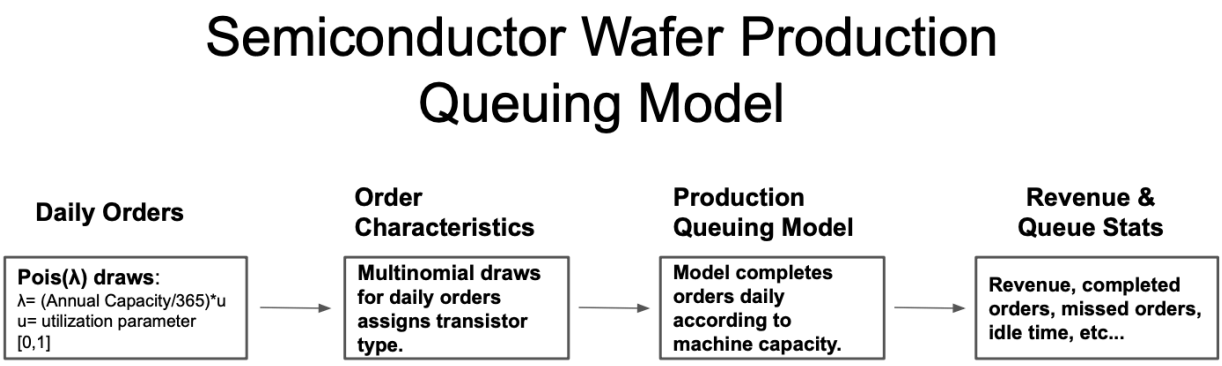

Figure 1: Basic flow chart for proposed queuing model.

The model runs over time tmax. For each cycle, it estimates daily orders using a poisson distribution with parameter λ. λ represents average daily capacity times a parameter u adjusting for fab's at less than full capacity. The model assigns each order to a specific machine's queue utilizing a multinomial distribution constructed from uniform [0,1] draws with probability pi bins for each chip type. The probability that a particular order is for a specific chip type was calculated using Figure 2 below:

---

[16] Vatas at 16-17.

[17] Vatas at 17-19.

[18] Vatas at 17-19.

Probabilities Used for Multinomial Model

$$p_i = \frac{\frac{c_i}{\pi_i}}{\sum_{i=1}^{i}(\frac{c_i}{\pi_i})}$$

c: annual chip type revenue

π: average chip type price

Figure 2: pi is annual revenue divided by sales price and normalized. Revenue and sales price estimates came from an annual report from a major semiconductor fabricator, Taiwan Semiconductor Manufacturing Company (TSMC), and price estimates from a recent Georgetown report called "AI Chips: What They Are and Why They Matter." [19]

Each machine's daily wafer production capacity is set by a vector of values. Each value represents all available machine capacity for the production of a specific transistor type. With each cycle, our model removes completed orders from each machine's queue based on that machine's available production capacity. If orders in the queue don't utilize a machine's full capacity, unused time is tracked as idle time. The model also allows caps to be placed on the total number of orders accepted in each machine's queue. Rejected orders are stored as lost orders alongside any order left in the queue at the end of tmax. Finally, wait times are calculated for each machine i at each time step by subtracting the time of the most distant order processed by machine i that day from the time step value.

Section 3: Results

To test the model, it was run multiple times with 9 different machine type queues. Machines produced 5nm, 7nm, 10nm, 16nm, 20nm, 28nm, 40/45nm, 65nm, and 90nm wafers at various rates. Two types of queue caps were tested. High and low. Queue caps for each machine type were 2,3,2,2,100,100,100,100, and 100 in the previously stated order. The model also tested a range of capacities. To check stability, the model was run with 5,10,15,30,50,70,60,100, and 500 as capacity parameters in the previously stated order. The model was run for 10 days. Despite the large range of parameters, the model operated as expected.

The model results tell us that despite relatively small production quantities the 5nm makes up a significant portion of revenue. Moreover, our results suggest we should review our cap policy and/or increase 7nm and 16 nm chip production since many orders were missed. In particular, the 16 nm deserves more attention, given it made up a significant portion of our revenue in spite of missing many orders. Finally, the model shows us that significant demand still exists for certain larger/older transistors like 28nm, 40/45nm, and 65nm.

---

[19] Kahn at 44; TSCM 2020.

	Model Run 1 (10 days)	Model Run 2 (10 days)	Model Run 3 (10 days)
Total Orders	3283	3337	3306
Wait Time	0.007	0.007	0.006
Machine Idle Time	11.7	9.6	11.14
Incomplete Orders	16	17	15
Projected Revenue	$8,084,189.00	$8,185,149.00	$8,084,189.00

Table 1: Table shows overall results from 3 queuing model runs with the aforementioned parameters.

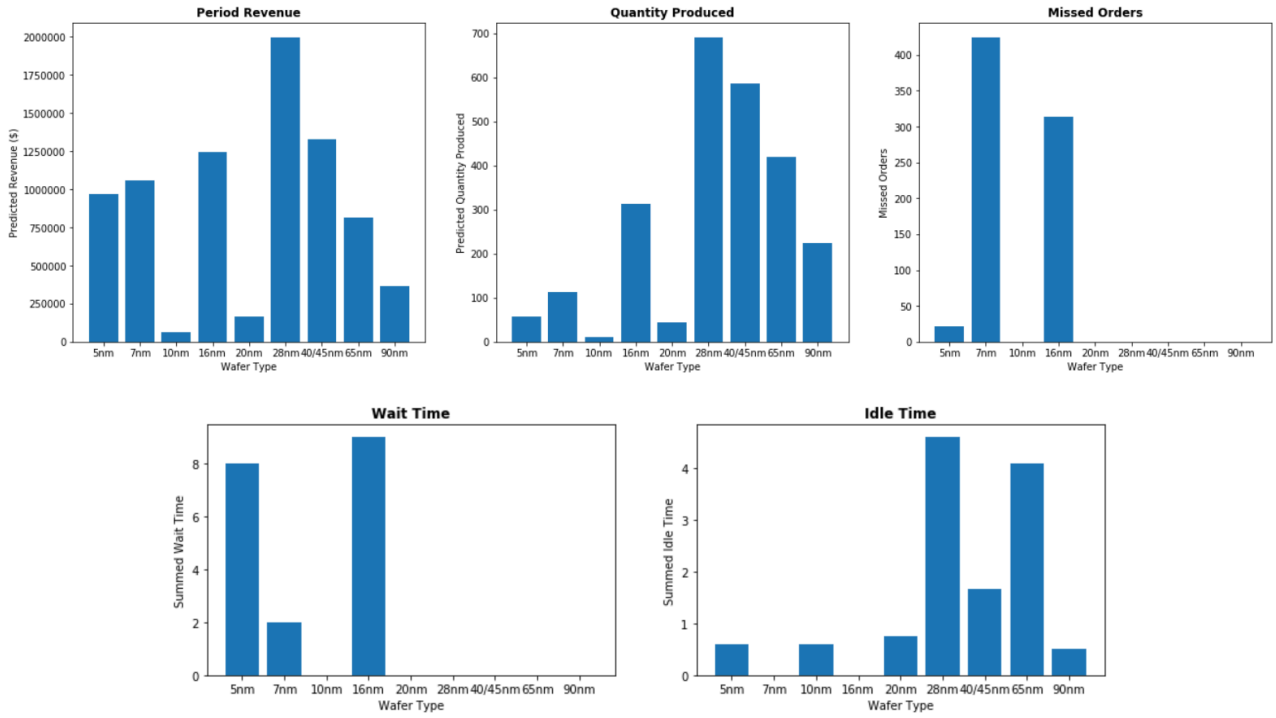

Figure 3: Plots show predictions from a single model run broken down by machine type.

Section 4: Discussion

The model has a number of caveats. In reality, order probabilities are likely time dependent and changing. Demand for specific transistor types changes at a rapid rate. For example, chip makers previously relying on 7nm technology will soon be transitioning to 5nm technology.[20] The model won't specifically account for these shifts in demand. A user must use domain knowledge to do

[20] Zafar.

so. The model assumes a machine's production capacity is relatively stable and can be represented as a scalar value. The model potentially over simplifies the process of scaling up production. As discussed, adding machine capacity can require a significant long term investment in labor and capital. The model's reliance on annual averages removes near term seasonal variation from order predictions. This means the model will operate best if run for a year smoothing out seasonal effects. Otherwise, a user assumes fabs experience fairly constant demand for different chip types over time. Finally, poisson draws assume variance is equal to λ. If we want to set a specific variance, the model can be adapted to use rounded draws from a normal distribution.

Section 5: Conclusion

The queuing model is a flexible and fairly straight forward method for estimating the impact of adding different allocations of production capacity to an existing fab. Since revenue and quantity is calculated for each machine, the model provides users the opportunity to design machine specific cost functions and incorporate more granularity into predictions. Cost functions vary by design type and might include inputs like labor per wafer, die per wafer, transistor costs, research and development, electricity, and capital investment depreciation. Cost per printed layer is relatable to a fab's weekly layer capacity and wafer size.[21] Different chip types may require different numbers of layers to complete a wafer. Large fab clusters can reduce costs. A 300 diameter mm wafer yields 2.4 times the IC's of a 200 mm wafer decreasing processing waste.[22] In some circumstances, fabrication can take up to 2 months, including hundreds of photographic and chemical processing steps.[23] Focusing on production by machine type accommodates these types of specifications. If provided a suitable inputs for a cost function and more definitive pricing information, revenue projections can estimate profits for different strategies. Moreover, the model provides a number of useful estimates, including information on what chip types are no longer performing well and where idle time exists.

Keeping the limitations in mind, the proposed model can identify important considerations for strategically increasing capacity. For example, model results may inspire assigning labor from machines with significant predicted idle time to machines not meeting demand. Economists believe there are optimal combinations of capital and labor that will maximize output by increasing specialization.[24] Finally, if investments in machinery or fab construction is required, the model provides a good starting point for analyzing the potential benefits of different types of production capacity.

---

[21] Bauer at 8.

[22] Platzer at 5.

[23] Platzer at 15.

[24] Besanko.

References:

1) Platzer, Michaela et al. 2020. *Semiconductors: U.S. Industry, Global Competition, and Federal Policy*. [online] Available at: <https://fas.org/sgp/crs/misc/R46581.pdf> [Accessed 4 May 2021].
2) Besanko. 2004. *Microeconomics Chpt 6: Inputs and Production Functions*. [online] Available at: <https://www.fep.up.pt/docentes/pcosme/micro-g102/Besanko-ch06.pdf> [Accessed 4 May 2021].

3) Jeong, E et al. 2021. *Why the Chip Shortage Is So Hard to Overcome.* [online] WSJ. Available at: <https://www.wsj.com/articles/why-the-chip-shortage-is-so-hard-to-overcome-11618844905> [Accessed 4 May 2021].

4) VerWey, J. 2019. *The Health and Competitiveness of the U.S. Semiconductor Manufacturing Equipment Industry.* [online] USITC. Available at: <https://www.usitc.gov/publications/332/working_papers/id_058_the_health_and_competitiveness_of_the_sme_industry_final_070219checked.pdf> [Accessed 4 May 2021].

5) Barbe, A et al. 2018. *Trade and Labor in the U.S. Semiconductor Industry.* [online] USITC. Available at: <https://www.usitc.gov/publications/332/journals/barbe_kim_and_riker_-_trade_and_labor_in_the_us_semiconductor_industry_.pdf> [Accessed 4 May 2021].

6) Bauer, H., 2021. *Semiconductor design and manufacturing: Achieving leading-edge capabilities*. McKinsey & Company. [online] Available at: <https://www.mckinsey.com/industries/advanced-electronics/our-insights/semiconductor-design-and-manufacturing-achieving-leading-edge-capabilities> [Accessed 4 May 2021].

7) Vatas, A. et al. 2020. *Government Incentives and US Competitiveness in Semiconductor Manufacturing*. [online] Available at: <https://www.semiconductors.org/wp-content/uploads/2020/09/Government-Incentives-and-US-Competitiveness-in-Semiconductor-Manufacturing-Sep-2020.pdf> [Accessed 4 May 2021].

8) McClean, E. 2018. *Advanced Technology Key to Strong Foundry Revenue per Wafer.* [online] IC Insights. Available at: <https://www.icinsights.com/data/articles/documents/1108.pdf> [Accessed 4 May 2021].

9) Kahn, S et al. 2020. *AI Chips: What They Are and Why They Matter.* [online] Georgetown Center for Security and Emerging Technology. Available at: <https://cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter> [Accessed 4 May 2021].

10) McClean, E. 2018. Revenue per Wafer Climbs As Demand Surges for 5nm/7nm IC Processes. [online] IC Insights. Available at: https://www.icinsights.com/data/articles/documents/1350.pdf [Accessed 4 May 2021].

11) TSMC 2020. TSMC Annual Report. [online] TSMC. Available at: <ttps://investor.tsmc.com/sites/ir/annual-report/2020/2020Annual%20Report_E_%20.pdf> [Accessed 5 May 2021].

12) Zafar, R. 2019. Apple A13 & Beyond: How Transistor Count And Costs Will Go Up. [online] wccftech. Available at: <https://wccftech.com/apple-5nm-3nm-cost-transistors/> [Accessed 4 May 2021].

13) Su, Jeff. April 2021. TSMC Quarterly Management Report. [online] TSMC. Available at: <https://investor.tsmc.com/english/encrypt/files/encrypt_file/reports/2021-04/c3503e3be4f98cd83817a8abff6e4586deaa3904/1Q21ManagementReport.pdf> [Accessed 5 May 2021].

14) TSMC. 2020. TSMC Annual Report. [online] TSMC. Available at: <ttps://investor.tsmc.com/sites/ir/annual-report/2020/2020Annual%20Report_E_%20.pdf> [Accessed 5 May 2021].