Transistors All the Way Down: Viability of Direct Volume Measurement (and Price Indexes) for Semiconductors[1]

David Byrne, Adrian Hamins-Puertolas, and Molly Harnish

January 6, 2024

[1]Prepared for presentation at the "Innovations in Economic Measurement" session at the 2024 ASSA Annual Meeting. The views expressed here are not represented to be the views of the Federal Reserve Board of Governors. The authors are grateful to Neil Thompson for feedback. The authors, listed alphabetically, contributed equally to this research. Andrea Garcia assisted with data curation and software. Please do not circulate these slides without consulting the authors.

Table of Contents

Overview: Motivation

Motivation

Solid-state electronics (semiconductors) is central to post-war macroeconomics:

- R&D and innovation
- productivity and growth
- trade and supply chains

Objective

Consistent output and prices:

- a full history
- for all products
- and all countries

"[Science] pronounces only on whatever, at the time, appears to have been scientifically ascertained, which is a small island in an ocean of nescience." – Bertrand Russell

Overview: Approach

Method

- direct volume measurement
- volume $\propto$ transistors
- implicit price indexes

Data requirements

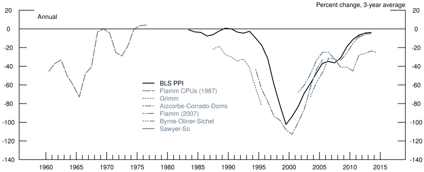
- plant-level capacity, product mix, and technology
- utilization rate
- nominal output

Today, we have results for microprocessors, memory, and other logic

Motivation: Time-series Inconsistency

Splicing together disparate indexes undermines analysis of long-run trends.

Fig. 1: Processor Price Indexes



Note: Indexes differ with respect to price data source, methodology, and scope (see text).
Source: Bureau of Labor Statistics, Grimm (1998), Aizcorbe, Corrado, and Doms (2003), Flamm (2007), Holdway (2000), Byrne, Oliner, and Sichel (2018), Sawyer and So (2017).

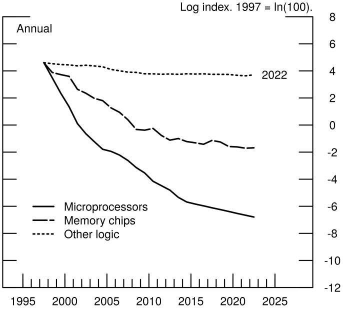
How to choose among and combine indexes that differ with respect to:

- methodology (matched-model vs. hedonic)
- data availability (model-level observations? market or list prices? item characteristics?)
- scope (all processor types? just U.S. production?)

Motivation: Cross-Product Inconsistency

Data limitations pollute cross-product comparisons

Fig. 2: FRB Price Indexes by Product



Note: MPU and memory price indexes are used in the industrial production index as shown. Other logic chip price index is adjusted for bias before use.
Source: Federal Reserve Board.

Are the true price trends this different or is this the result of sources and methods?

- MPUs and memory:
  detailed data and solid methodology
- Other logic:
  index of average prices for coarsely defined chip types.

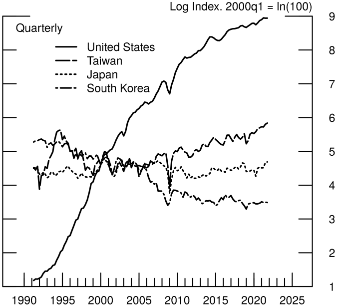
Are "other logic" prices this stagnant? These chips include:

- hardware accelerators for graphics (GPUs), AI (TPUs), signal processing (DSPs), etc.
- custom processors for cell phones and other devices (ASICs).

Motivation: Cross-Country Inconsistency

Cross-country comparisons lead to spurious results.

Fig. 3: Implied Chip Quality by Country



Note: Natural log of industrial production index divided by (eight-inch equivalent) wafers shipped.
Source: National statistical agencies, SEMI, Gartner, Inc., authors' calculations.

A coarse indicator of “quality” for national chip industries: real output divided by silicon wafers shipped

Cross-country quality comparisons are counter-intuitive:

- U.S. quality outpacing Taiwan?
- Outright quality decline in Korea?
- There is some room for variation due to product mix, but not this much.

Table of Contents

Chip Manufacturing (top down)

Chip technology is characterized by a small set of parameters, plus "cleverness."

Production costs fall with:

- increasing die (chip) size
- miniaturization (dimension reduction)
- better circuit layout ("cleverness")

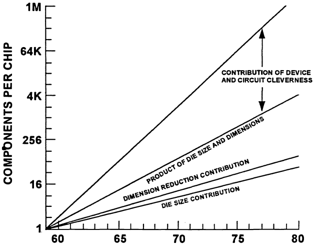


Figure 8. Resolution of the increase in complexity into die size, dimension reduction and "cleverness" factors. Proceedings IEEE IEDM 1975.

Fig. 4: Wafer & Die Size

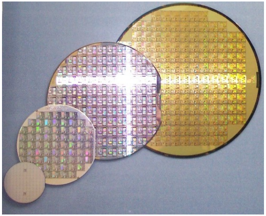
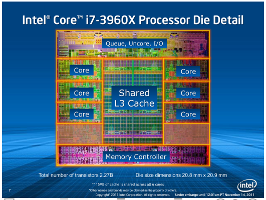


Fig. 5: Layout

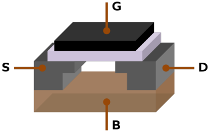


Chip Manufacturing (bottom up)

Transistors are the key components of (most) semiconductors.

- Transistors are electronic on/off switches.
- Gates are logical operators built from transistors.
- Integrated circuits are built from gates.

Fig. 6: Transistor



G: gate, B: body, S: source, D: drain, Pink: insulation.

Fig. 7: "NAND" Gate

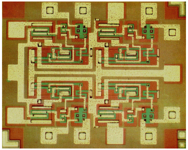


Table of Contents

Product Quality: Hedonic Approach

Hedonic Analysis

- Can be seen as grounded in consumer theory (e.g. Lancaster) where utility is defined over a set of primitives and goods are characterized as bundled quantities of those desiderata.
- In principle, one could leverage this structure to construct a cost-of-living index.
    - Erickson and Pakes; Anderson, de Palma, Thisse; and descendants.
- However, the characteristics corresponding to these primitives are often unknown, unobserved, or both; a supply side is needed as well; and myriad functional form assumptions are needed for identification.
- Typically, hedonic price indexes simply find a "reduced-form" correlation between observed characteristics and observed average prices.
- Hedonic regression is not a panacea. Suitability depends on market structure and data available.

Product Quality: Matched-Model Approach

Quality-Adjustment in Matched-Model Indexes

- Prices for goods in a representative basket are observed repeatedly over time.
- Item-level price changes (relative to the previous period) are averaged across the basket using weights reflecting their relative importance
- Assuming (1) the quality of each item does not change over time and (2) the law of one price holds, incumbent item prices fall when new, lower (quality-adjusted) priced items enter the market . . .
- . . . the workhorse matched-model index controls for quality change.

Product Quality: Benchmark Approach

Benchmark Approach

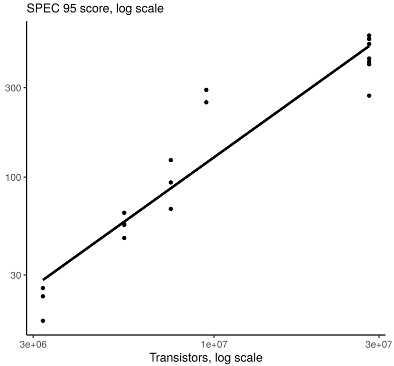
- Hedonics requires the analyst to become an expert on the product. An alternative: Ask existing experts.
- For benchmarks, experts develop tests that mimic common use cases (either complete programs or fundamental computing tasks).
  - search, encryption (MPUs); graphics rendering (GPUs); system simulation (both MPUs and GPUs)
- Benchmark scores are aggregates of results from these tests.
- These scores can be treated as direct measures of quality, which yield constant-quality price indexes when divided into nominal sales.

Unfortunately, benchmarks are not available for all chip types in all time periods . . .

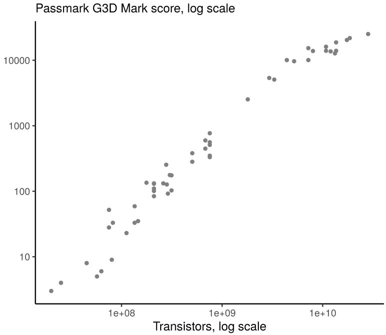
Product Quality: Benchmarks (continued)

...but benchmarks are are highly correlated with transistor counts

MPUs (SPEC 1995)



GPUs (Passmark)



Note: Nvidia GeForce GPUs released from 1999 to 2022.
Source: Wikipedia; Passmark.

Product Quality: Benchmarks (continued)

An appeal to authority: The views of Zvi Griliches and Jack Triplett on benchmarks

- Griliches and Ohta (1976): "ideally, quality adjustments should be based on performance variables, which presumably enter the utility function directly, not physical characteristics"
- Triplett (2005): "Benchmark measures have the advantage that they measure machine performance, rather than measuring some proxy for machine performance, or some input that may influence machine performance"

Table of Contents

Index Construction: Intuition

Index intuition

- real output $\propto$ performance $\propto$ transistors
- $\implies$ we use (adjusted) transistor counts to measure real output
- (and divide real into nominal output to get price indexes)

Le mieux est l'ennemi du bien.

– Voltaire, 1770.

Index Construction: Details

Real output Q of product p for plant i in period t

Wafers per plant: $N * U$

- plant capacity: $N_{i,t}$ (SEMI, Inc.)
- utilization: $U_{i,t}$ (Census Bureau)

Chips per wafer: Δ

- wafer diameter: $D_{i,t}$ (SEMI, Inc.)
- die area: $S_{i,t} = f_S(p_{i,t}, t)$
  (International Technology Roadmap for Semiconductors)
- chips per wafer: $\Delta_{i,t} = \frac{\pi D_{i,t}^2}{4S_{i,t}} - \frac{\pi D_{i,t}}{\sqrt{2S_{i,t}}}$ (DeVries, 2005)

Transistors per chip:

- geometry/line width/node: $G_{i,t}$
- chip layout factor: $L_{i,t} = f_L(p_{i,t}, t)$

Suppressing plant, product, and time subscripts, each plant produces

$$Q = (S * G^{-2} * L) * (\Delta) * (N * U) \tag{1}$$

Index Construction: Caveats

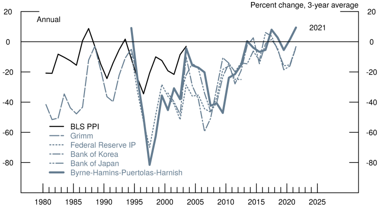
Caveats

- Other aspects of performance, notably energy efficiency, are not captured.
- Additional heterogeneity within our coarse product categories.
- The relationship between transistors and performance may change.

Table of Contents

Results: Memory Price Index

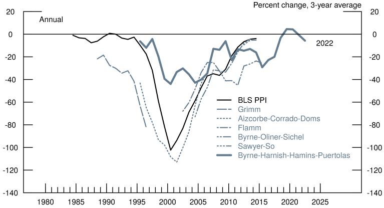
Fig. 8: Memory Price Indexes



Note: Indexes differ with respect to price data source, methodology, and scope (see text).
Source: Bureau of Labor Statistics, Federal Reserve Board, Grimm (1998), Bank of Korea, Bank of Japan.

- Our index (thick grey line) aligns with official indexes.
- This is good news. When chips are logically simple and data is abundant, we have confidence in the official index.
- Recent divergence is likely the result of incomplete adjustment for memory cell stacking in recent chips.

Fig. 9: MPU Price Index

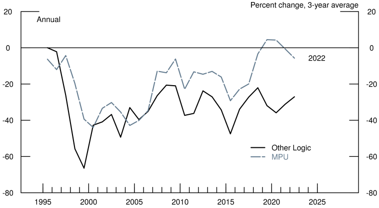


Note: Indexes differ with respect to price data source, methodology, and scope (see text).
Source: Bureau of Labor Statistics, Grimm (1998), Aizcorbe, Corrado, and Doms (2003), Flamm (2007), Holdway (2000), Byrne, Oliner, and Sichel (2018), Sawyer and So (2017).

- Our index accelerates (falls faster) in the late 1990s, as do others.
- But it accelerates by far less. The MPU contribution to the IT boom is smaller than previously thought.
- MPU price declines slow recently, suggesting weaker contribution of IT hardware to growth.

Results: Other Logic Price Index

Fig. 10: Other Logic v. MPU Prices



Note: Other logic includes hardware accelerators (e.g. GPUs), ASICs (application specific integrated circuits), and miscellaneous other non-memory, non-MPU logic chips.
Source: Authors' calculations. (See text for details.)

- Semiconductor innovation has not slowed down, but has shifted. Prices for other logic fall faster than MPU prices!
- This accords with expectations. Derived demand for GPUs from crypto, AI drives innovation.

Takeaways

Takeaways

- Feasible direct volume measures for output and prices.
- Consistent across time, countries, and products.
- Effort and industry expertise required are modest.
- Data requirements are easily met:
  plant technical detail, utilization, and nominal output.
- Results shed new light on key macro questions.

Thank you

disconnected measurement $\implies$ *disconnected macro.*

We look forward to your feedback.