Transistor Count: A Flawed Metric

In the tech industry, transistor count and transistor density are often portrayed as technical achievements and milestones. Upon the release of a new processor or SoC, many a vendor brags about the complexity of their design, as measured by transistor count. As a recent example, when Apple released the A13 Bionic inside the iPhone 11 generation, the company crowed that it contains 8.5 billion transistors, and in 2006, Intel similarly bragged about Montecito, the first billion-transistor processor.

For the most part, these constantly increasing transistor counts are a consequence of Moore's Law and the drive to ever greater levels of miniaturization. As the industry moves to newer process technologies, the number of transistors per unit area keeps on rising. For this reason, transistor count is often considered a proxy for the health of Moore's Law, although that is not quite fully accurate. Moore's Law in its original form observes that the transistor count of an economically optimal (i.e., minimum cost per transistors) design doubles every two years. But from a customer standpoint, Moore's Law is really a promise that the processors of tomorrow will be even better and more valuable than the processors of today.

In reality, transistor density varies considerably based on the type of chip and especially the type of circuitry within the chip. Worse yet, there is no standard way of counting transistors and the numbers can vary by 33-37% for the same design. The net result is that transistor count and density are only approximate metrics and focusing on those particular numbers risks losing sight of the bigger picture.

Product Objectives Influence Design Style

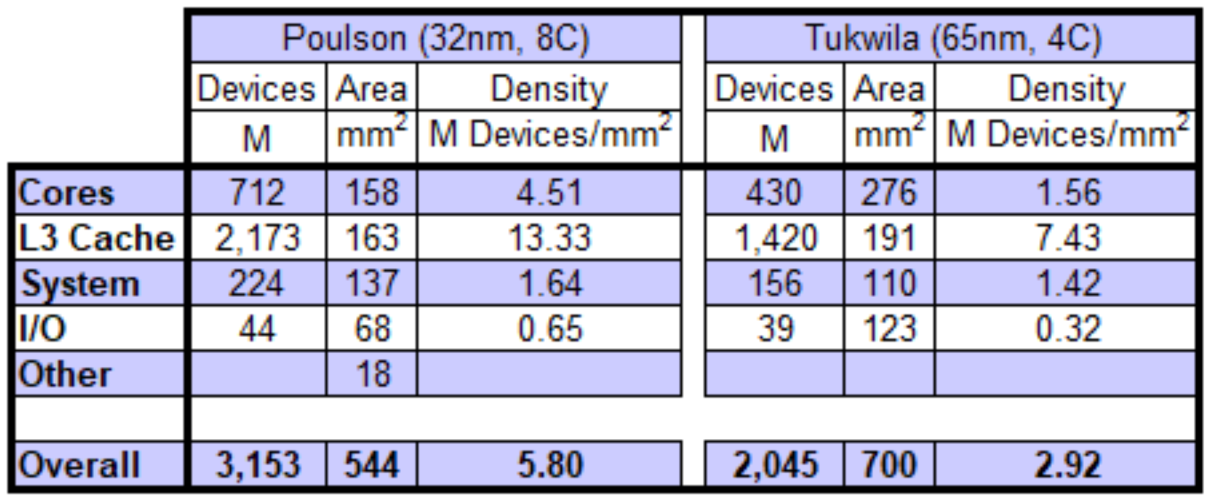
The transistor density is intimately related to the overall objectives and design style. Comparing substantially different designs such as a fixed-performance ASIC (e.g., Broadcom's Tomahawk 4 25.6Tb/s switch chip or Cisco's Silicon One 10.8Tb/s router chip) and a high-performance datacenter processor (e.g., Intel Cascade Lake or Google's TPU3) is misleading at best.

An ASIC needs to deliver the targeted throughput, but does not benefit from any incremental frequency. For example, the Cisco Silicon One is intended for high-speed networking using 400Gbps Ethernet and there is no advantage from boosting the frequency by 10%; 400Gbps is the standard set by IEEE, and the next step after that is 800Gbps. As a result, most ASIC design teams will tend to optimize for minimum cost with highly automated design tools, fewer custom circuits, and dense transistors.

In contrast, a faster server chip can usually command higher prices and therefore will always benefit from any incremental frequency. For example, the Xeon 8268 and 8260 are both 24-core parts and the main difference is the base frequency (2.9GHz and 2.4GHz), which translates into about $1,600 difference in list price. The server design team will therefore optimize for frequency. High-speed designs like the server processor tend to use more custom circuit design and larger transistors that have greater drive strength and reduced variability. In modern FinFET-based designs, this translates into more transistors with 2 fins, 3 fins, or even more. In contrast, lower-speed logic like an explicitly parallel GPU or ASICs often employ the densest transistors that use just a single fin, sacrificing clock speed to improve density. Similar to high-speed logic, ultra-low leakage transistors are often larger as well.

Transistor Count and Density are Determined by Design Balance

An even bigger influence on transistor count and density is the actual composition of the chip. Every modern design is built from some combination of logic for computation, memory (typically SRAM) for storage, and I/O for communication. However, these three constituents are all differ radically in terms of density, as illustrated in Table 1. Poulson and Tukwila are platform compatible and share the same overall goals of delivering high-performance and the highest levels of reliability for mission critical servers.



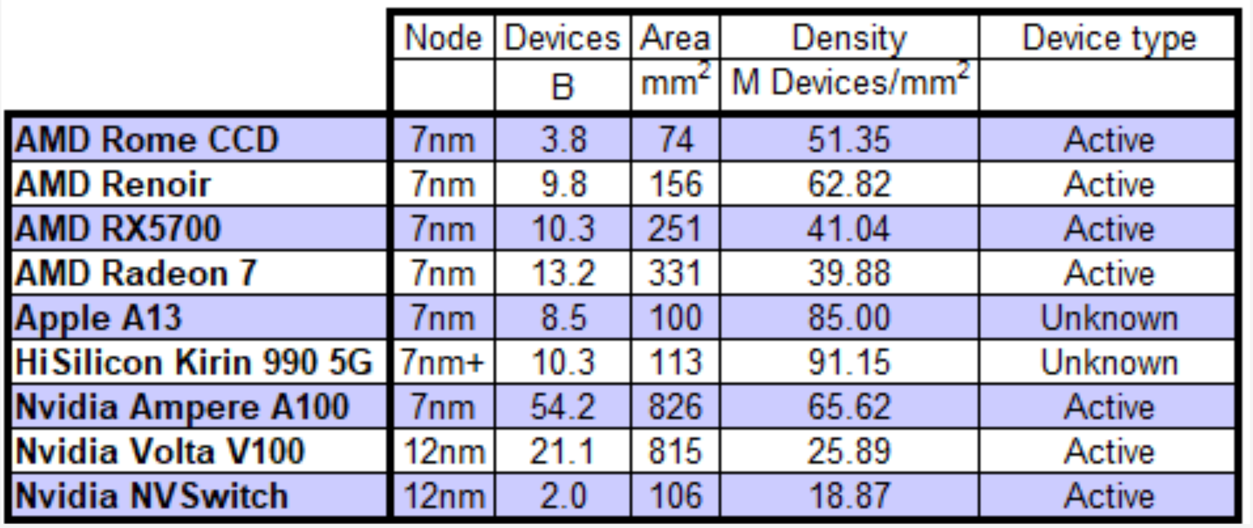
	Poulson (32nm, 8C)			Tukwila (65nm, 4C)		
	Devices	Area	Density	Devices	Area	Density
	M	mm^2	M Devices/mm^2	M	mm^2	M Devices/mm^2
Cores	712	158	4.51	430	276	1.56
L3 Cache	2,173	163	13.33	1,420	191	7.43
System	224	137	1.64	156	110	1.42
I/O	44	68	0.65	39	123	0.32
Other		18				
Overall	**3,153**	**544**	**5.80**	**2,045**	**700**	**2.92**

Table 1. Transistor count and density for major regions for the Poulson and Tukwila generations of Itanium processors

The processors comprise four major regions: CPU cores, L3 cache, the system interface, and I/O. Based on the reported information, Poulson also includes 18mm^2 of die area for whitespace or other functions. The CPU core region includes the cores and performance optimized L1 and L2 caches and is dominated by high-speed logic that targets operation over 1.7GHz for Tukwila and 2.5GHz for Poulson. The large L3 caches (24MB for Tukwila and 32MB for Poulson) are designed for maximum capacity and uses the densest 6-transistor (6T) SRAM cells possible with dedicated power rails to ensure stability. The system region includes an assortment of functions – a crossbar for communicating I/O and memory traffic across the die, QPI and memory controllers, home agents for the directory-based coherency protocol and directory caches, and power management units. The system region is generally not as dense because the logic is fixed-frequency and many of the biggest components (e.g., the crossbar) are dominated by large high bandwidth busses crossing the die, rather than transistors. Lastly, the I/O region contains the physical interfaces for external communication, which are implemented using high-speed serial interconnects including four full-width QPI links, two half-width QPI links, and two FB-DIMM2 or Scalable Memory Interfaces (SMI) that fan out to four channels of memory. The interconnects use differential signaling and total around 600 pins.

Quantitatively, these two processors illustrate crucial trends that hold true across all major chip designs. First of all, the variation in transistor density between different regions of the chip is enormous – over 20X, and dwarfs the factor of 2X that is associated with a single generation improvement according to Moore's Law. Naturally, the cache region which primarily comprises ultra-dense SRAM is the densest and makes up most of the transistors in each design. The cache is about 3-5X as dense as the computational logic in the cores, again larger than a single scaling factor. The I/O is the least dense portion of the two designs, because it contains many delicate analog circuits such as PLLs and DLLs, digital filters, and the large, high-voltage I/O transistors that are used to transmit and receive off-chip data. Additionally, many I/O regions must occupy enough of the edges of the chip to connect all the pins and the area is determined by the number of pins, not the density of the circuits.

The data above clearly demonstrates that the transistor density of modern chips is strongly a function of the purpose and composition of the chip. To take an extreme example, imagine a 32nm design that is based on Poulson, but with no L3 cache – it would have a transistor density of around 2.57M/mm^2, well under half the density of the actual Poulson design. In the other direction, a hypothetical version of Poulson with just compute and cache and no I/O or system regions would have a transistor density of 9M/mm^2.



	Node	Devices	Area	Density	Device type
		B	mm^2	M Devices/mm^2	
AMD Rome CCD	7nm	3.8	74	51.35	Active
AMD Renoir	7nm	9.8	156	62.82	Active
AMD RX5700	7nm	10.3	251	41.04	Active
AMD Radeon 7	7nm	13.2	331	39.88	Active
Apple A13	7nm	8.5	100	85.00	Unknown
HiSilicon Kirin 990 5G	7nm+	10.3	113	91.15	Unknown
Nvidia Ampere A100	7nm	54.2	826	65.62	Active
Nvidia Volta V100	12nm	21.1	815	25.89	Active
Nvidia NVSwitch	12nm	2.0	106	18.87	Active

Table 2. Transistor count and density for selected 7nm and 12nm chips, as reported by vendors.

Table 2 contains details on several chips manufactured on TSMC's 12nm and 7nm process nodes that highlight the impact of design composition on density. As a first illustration, AMD's Radeon VII and RX 5700 are relatively similar GPU designs on the same node and have nearly identical density. On the other hand, AMD's Renoir and Nvidia's A100 are about 1.5X the density of these GPUs, perhaps reflecting a focus on density, or potentially more mature design tools. Another useful comparison is Nvidia's V100 GPU and the NVSwitch, which is an 18-port NVLink switch. They are on the same node, but the latter is primarily I/O and on-die routing for NVLink and as a consequence, the V100 is 1.37X denser than the NVSwitch.

Lastly, the two smartphone SoCs are 1.35X-2.29X denser than the rest of the 7nm processors. This impressive density is possibly due to the different optimization targets – smartphone SoCs are tailored for low-cost and high-density, while AMD's processors tend to target high performance. Additionally, Apple and HiSilicon are larger and more profitable companies that can afford larger design teams and greater optimization efforts. However, it is also possible that the transistor counts and density for the mobile SoCs are the resulting of a different form of transistor accounting. The last column in Table 2 indicates how the vendor is counting transistors, which we will discuss in greater detail on the next page.

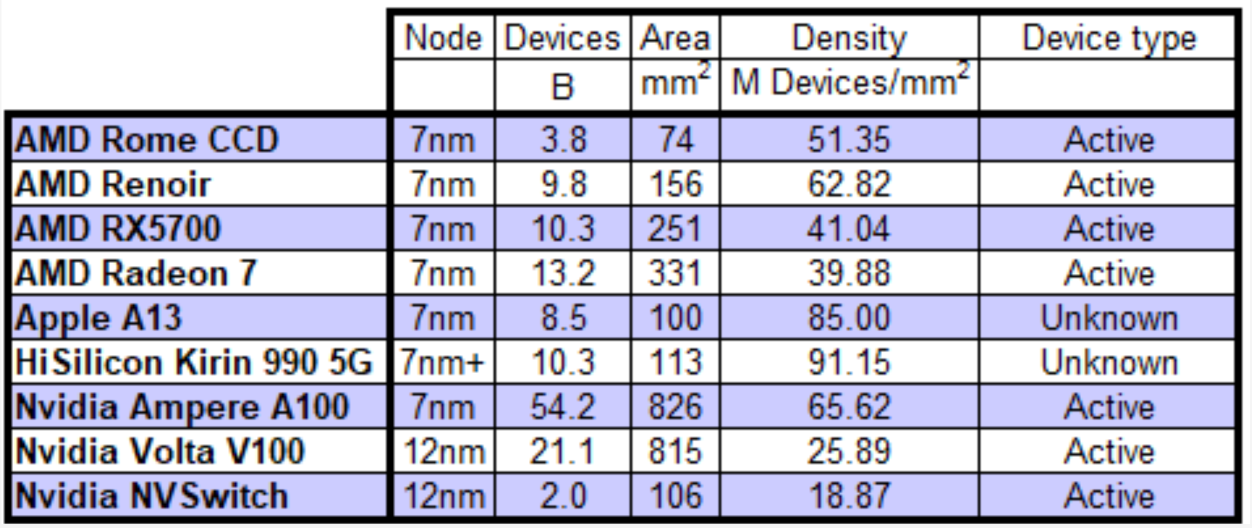
Not All Transistors are Equal

Another challenge to using transistor count or density as a metric is that it is ambiguous and potentially misleading. Typically, we think of transistors as the physical implementation of logical blocks and circuits. For computation, this could be anything as large as CPU core or floating-point unit to something as small as an inverter. For storage, this might be a cache, a register file, a content-addressable-memory (CAM), or an SRAM bit-cell. For analog or I/O, this could be a PLL or an off-chip transmitter or receiver. The transistors that physically implement these blocks are referred to as active transistors (which are distinct from schematic transistors). In reality though, not all transistors are created equal and modern chips are manufactured with many transistors that are not active. The transistors formed during the manufacturing process are descriptively known as layout transistors. The layout transistors include the active transistors, as discussed above, but also dummy transistors and transistors used as decoupling capacitors.

Dummy transistors are inserted into a design to improve yield. For example, certain annealing and etching steps in the manufacturing process work best on a relatively uniform surface, and inserting extra transistors into empty areas improves the uniformity and therefore the yield. For many analog circuits, these extra transistors are necessary to achieve the desired performance. As another example, modern FinFET performance varies based on the stress on the transistor, which is a function of the other nearby transistors. Achieving the right performance may require placing transistors nearby to obtain the right stress.

While dummy transistors are commonly used, they are not particularly numerous. In contrast, decoupling capacitors built from MOSFETs (or decap cells) are used extensively. Generally speaking, the logic in modern chip designs never achieves 100% areal efficiency. For all the marvel of modern design tools, there is typically empty whitespace between individual logic cells (e.g., a NAND gate), between functional units (e.g., the L1D cache), and between entire IP blocks (e.g., a CPU core). Whitespace is a consequence of the tools struggling to follow design rules that ensure yield and frequency, use the available resources (such as routing layers), and piece together an electrical engineering jigsaw puzzle of logic cells, functional units, and blocks. Whitespace can account for 10-25% of a design. To ensure yield, the die must be relatively uniform and the whitespace cannot be truly empty. Many designs will fill the whitespace with decap cells to provide decoupling capacitance for power delivery and thereby improve operating frequency. In addition, some designs will place decaps within standard cell libraries. Decap transistors are the dominant source of non-active layout transistors, but hard data is difficult to obtain.

Our friends at TechInsights perform circuit-level analysis that includes the number of active and layout transistors for small portions of a die. They were kind enough to share some of these analyses for a handful of 7nm SoCs. The data is based on a small number of sample locations within each SoC, typically the GPU, which will have the greatest transistor density. They found that in the small sampled regions that the active transistors were between 70-80% of the total, and the remaining 20-30% of layout transistors were decap and dummy devices. These numbers are based on limited samples, because this analysis is fairly expensive and time-consuming. To confirm and elaborate, we gathered numbers on several modern designs and found that active transistors are commonly 63-66% of the layout transistors and that 33-37% of the layout transistors are decap cells. The TechInsights numbers are probably low because they are primarily looking at the densest logical portions of a SoC, rather than including whitespace which would include more decap transistors.



	Node	Devices	Area	Density	Device type
		B	mm^2	M Devices/mm^2	
AMD Rome CCD	7nm	3.8	74	51.35	Active
AMD Renoir	7nm	9.8	156	62.82	Active
AMD RX5700	7nm	10.3	251	41.04	Active
AMD Radeon 7	7nm	13.2	331	39.88	Active
Apple A13	7nm	8.5	100	85.00	Unknown
HiSilicon Kirin 990 5G	7nm+	10.3	113	91.15	Unknown
Nvidia Ampere A100	7nm	54.2	826	65.62	Active
Nvidia Volta V100	12nm	21.1	815	25.89	Active
Nvidia NVSwitch	12nm	2.0	106	18.87	Active

Table 2. Transistor count and density for selected 7nm and 12nm chips, as reported by vendors.

The data makes it abundantly clear that there is often a big difference between the number of active and layout transistors in a chip. Unfortunately, many companies do not specify which number they are quoting. The data on AMD and Nvidia processors from Table 2 are all from technical papers. Based on discussion with these two vendors, the numbers are active transistors as described in the last column. Based on some informal discussions, it appears that the HiSilicon Kirin 990 5G number may actually be layout transistors, which would help explain the discrepancy between these designs. It is unclear whether Apple’s A13 is implemented using 8.5 billion active transistors or layout transistors. The former would be an impressive achievement in density.

It doesn’t seem reasonable to count these dummy and decap transistors in the same way as active transistors. Active transistors implement the functions and features that customers value, whether it is CPU cores, power gating to improve idle power, neural network accelerators, or cache. However, decap and dummy transistors are overhead and don’t directly add value, and in some cases are worse than more sophisticated technologies. For example, IBM’s deep trench capacitors are far superior to decoupling capacitors and enable building dense eDRAM and reducing system cost. Similarly, Intel’s FIVR boosts platform efficiency and relies on MIM capacitors and virtually eliminates the need for decoupling capacitors in the package and board and likely reduces on-die decap transistors as well. In both cases, reducing the number of decap transistors is a benefit. The central point of Moore’s Law is creating value for customers by using additional active transistors productively, and decap and dummy transistors don’t really contribute to those goals.

It’s Not How Many Transistors, But How You Use Them

Putting this all together, it is clear that transistor count and transistor density are highly problematic metrics. Both are strongly influenced by the overall design and the ratio of critical blocks like computational logic, SRAM, and I/O. Of the three, SRAM is by far the densest, so a modest change in the size of a cache could produce big changes in transistor count with a minimal impact on performance and value. Moreover, not all layout transistors are equal. Active transistors are the fundamental building block for valuable components like CPUs and GPUs. On the other hand, decap and dummy transistors are more akin to overhead. Hopefully most companies are unlikely to conflate active and layout transistors, but it is important to distinguish the two when comparing designs.

Despite the problems with transistor count, it is potentially useful in a very coarse-grained fashion. It’s almost certainly true that a processor with 100-billion transistors is more complex and more valuable than one with 100-million. The analysis probably still holds for a 2X difference in transistor count – especially for something that is targeting explicitly parallel workloads like a GPU, or for two very similar processors (e.g., two smartphone SoCs or two server processors). But realistically, it’s hard to believe that a modest difference in transistor count necessarily translates into meaningful difference in value. In fact, AMD’s Radeon VII and RX 5700 provide a great counterpoint. The Radeon VII packs in 28% more transistors, but delivers fairly similar performance in part because the RX 5700 line uses a more advanced architecture. Moreover, the RX 5700 line is far less expensive since it uses GDDR6 instead of HBM2. When it comes to actual value to customers, it’s not about the transistor count, but how the transistors are used. Modest differences in transistor count just don’t matter compared to good architecture, feature selection, and other factors.

Many of these same criticisms hold for transistor density and process technology. If a modest increase in transistor count doesn’t meaningfully impact customer value, then it seems unlikely that a corresponding modest increase in density would tip the scales either. On the other hand, factors such as transistor performance, dynamic power, idle power, toolchain and third party IP support, wafer availability, and advanced features such as RF or optical devices can bring significant value to the table for the right products. Density is just one of many aspects of a process, and focusing on it too much risks missing the forest for the trees.