Contributing Author: David Floyer The latest Arm Neoverse announcement further cements our opinion that its architecture, business model and ecosystem execution are defining a new era of computing; and leaving Intel in its dust. We believe the company and its partners have at least a two year lead on Intel and are currently in a [...]

© 2021 Wikibon Research | Page 1
Contributing Author: David Floyer

The latest Arm Neoverse announcement further cements our opinion that its architecture, business model and ecosystem execution are defining a new era of computing; and leaving Intel in its dust. We believe the company and its partners have at least a two year lead on Intel and are currently in a far better position to capitalize on the major waves driving technology industry innovation. To compete, our view is Intel needs a new strategy (Gelsinger is bringing that), financial support from U.S. and EU governments, and, very importantly...volume for its foundry business. And that’s where Apple could be the key.

In this Breaking Analysis, we’ll explain why and how Apple could hold a key to saving Intel’s (and America’s) semiconductor industry leadership position. In addition, we’ll use ETR data on Dell’s laptop demand to explain why a booming business for Dell, may not be all that great for Intel. Finally, we’ll further explore our scenario on the evolution of computing and what will happen to Intel if it can’t catch up. Here’s a hint...It’s not pretty.

Assumptions Informing our Research Conclusions

Volume makes kings. Our most fundamental premise is that, all things being equal, volume is the most important determining factor to semiconductor leadership. It drives cost, experience, quality and innovation.

Our Fundamental Assumptions

- Arm wafer volume 10X that of x86 wafers
- Current chip shortage will last no less than 3 years
- Technically Moore’s Law is waning but its fundamental premise is accelerating...dramatically
- Arm is attacking the enterprise and cloud vendors (e.g. Graviton) are its entry point. If Intel can’t respond, Arm will gain 50% of enterprise semi spend by ‘30
- The cloud is expanding and a new model of computing is emerging, bringing new workloads, capabilities and value
- Geopolitics is a major driving force. The U.S. government will do/should do everything in its power to save Intel and the U.S. industry from irrelevance

We’ve pointed out many times that we believe Arm wafer volumes are approaching 10x those of x86 wafers. This means that manufacturers of Arm chips have a significant cost advantage over Intel. We’ve covered that extensively but we repeat it because when we see news reports and analysis in print it’s not a factor that is often highlighted. We believe this is the single most important issue that Intel faces and it’s why we feel Apple could be Intel’s savior. We’ll come back to that.

We project that the chip shortage will last no less than three years and perhaps longer.

As we reported in a recent Breaking Analysis, While Moore’s Law is waning, the outcome of Moore’s Law – i.e. the doubling of processor performance every 18-24 months, is accelerating. We have observed over the
past five years, and continue to project, a quadrupling of processor performance every two years.

Arm is attacking the enterprise and we see hyperscalers as the tip of their entry spear. AWS’ Graviton chip is the best example. Amazon and other cloud vendors that have engineering and software capabilities are making Arm-based chips capable of running general purpose applications. This is a huge threat to x86 and if Intel doesn’t respond quickly, we believe Arm will gain a 50% share of enterprise semiconductor spend by 2030.

We see the definition of cloud expanding. Cloud is no longer a remote set of services “out there in the cloud.” Rather it is expanding to the so-called edge, where the edge could be a datacenter, a data closet or a true edge device or system. And Arm is, by far in our view, the best positioned to support the new workloads and computing models that are emerging as a result.

Finally, geopolitical forces are at play. We believe the U.S. government will do, or at least should do, everything possible to ensure that Intel and the U.S. chip industry regain its leadership position in the semiconductor industry. If they don’t, the U.S. and Intel could fade to irrelevance.

**Gelsinger’s Geopolitical Trump Card**

Here’s a map of the South China Sea.

Way off in the pacific we’ve superimposed a little pie chart. We asked ourselves – if you had 100 points of strategic value to allocate. How much would you put in the semiconductor manufacturing bucket and how much would go to design? Our conclusion is 50/50.

Historically, because of Intel’s dominance with x86 volume, the United States was #1 in both strategic areas. But today, that orange slice of the pie is dominated by TSMC, thanks to Arm volumes. We’ve reported on this dynamic extensively and don’t want to dwell on the topic for too long. But on all accounts, cost, technology, volume...TSMC is the clear leader.

China’s President Xi has a stated goal of unifying Taiwan by China’s centennial in 2049. Will this tiny island nation, which dominates a critical part of the strategic semiconductor pie, go the way of Hong Kong and be
Will Taiwan go the way of Hong Kong and be subsumed into China? What are the implications for the U.S. semiconductor industry, Intel and Apple?

Admittedly, military experts say it would be very hard for China to take Taiwan by force without heavy losses and serious international (and domestic) repercussions. The U.S. military’s presence in the Philippines, Okinawa and Guam, combined with support from Japan and South Korea would make it even more difficult. And certainly the Taiwanese people, you would think, prefer their independence.

But Taiwanese leadership ebbs and flows between those hard liners pushing for independence and those more sympathetic to China. Could China, for example, use cyber warfare to, over time, control the narrative in Taiwan? The most embryonic state of the narrative is the meme. Control the meme you can control the narrative. Control the narrative, you can control the idea. If you can control the idea, you can control the belief system. If you can control the belief system, you can control the population, without firing a shot. So is it possible that over the next 25 years, China could weaponize propaganda to reach its objectives with Taiwan?

Maybe it’s a long shot but if you’re a senior strategist in the U.S. government (or Apple), would you want to leave that to chance? We wouldn’t.

Let’s park this for now and come back later to this dynamic.

“Moore’s Law” is Actually Accelerating

For at least four decades, the technology industry has marched to the cadence of Moore’s Law. While the strict definition of the principle is slowing, the rate of semiconductor performance gains is accelerating, dramatically.
As we first reported a few weeks ago, while Moore’s Law is moderating, the outlook for cheap, dense and efficient processing power has never been better.

While Moore’s Law is moderating, the outlook for cheap, dense and efficient processing power has never been better...Processing power is now quadrupling every 24 months versus the historical doubling.

The chart above shows two simple log lines. The blue line depicts traditional Moore’s Law progression and the orange shows the current pace of system performance improvement—measured in trillions of operations per second. The math on the historical annual rate of processor performance improvement with x86 comes out to around 40% per year. That rate is slowing...it’s now down to around 30% annually. So we’re not quite doubling every 24 months any more. And that’s why people say Moore’s Law is dead.

But if you look at the combinatorial effects of packaging CPUs, GPUs, NPUs, accelerators, DSPs and all the alternative processing power you find in SoCs, (system on chip), and eventually system on package, it’s growing at more than 100% per annum. This means processing power is now quadrupling every 24 months.

Arm has Reset Expectations for Moore’s Law

Arm has redefined the core processor model for a new era of computing. It made an announcement last week which recycled some content from last September but also put forth new proof points on adoption...
and performance.

The company laid out three components in its announcement: 1) Neoverse V1, which is all about extending vector performance. This is critical for high performance computing. Once thought a niche but it’s the AI platform. AI workloads are not a niche; 2) Arm announced the Neoverse N2 platform based on the recently introduced Armv9 - this got a performance boost of 40%; and 3) The CMN-700, which Arm said was the industry’s most advanced Mesh interconnect. This is the glue for the V1 and N2 platforms. The importance is it allows for more efficient use and sharing of memory resources across components of the system package...We talked about this extensively in previous episodes and will dig deeper into the importance of this technology later in the post.

This wheel diagram above underscores the completeness of the Arm platform. Arm’s approach is to enable flexibility across an open ecosystem, allowing for value add at many levels. Arm has built the architecture and design and facilitates standards across an open ecosystem that provides value added software to the package.

Importantly, Arm has created and is the steward of the standards and specifications by which software run on its platforms. It can, with certainty, certify that the foundry can make chips to a high standard; and importantly that all applications will run properly. In other words, if you design an application, it will work across the ecosystem and maintain backwards compatibility. Like Intel has done for years, but Arm as we’ll see next is positioning for not only existing workloads, but all the emerging and high growth applications.

Arm’s TAM Expansion Strategy

Below is the Arm total available market. Intel is following a similar path based on Gelsinger’s announcement earlier this year, but Arm and its ecosystem is well ahead in our opinion.
We think the end market spending value of just the chips going into these areas is $600B today and will grow to $1 Trillion by 2030. In other words we’re allocating the value of the end market spend in these sectors to the marked up value of the silicon as a percentage of the total spend. It’s enormous.

So the big areas are hyperscale clouds, which we think is around 20% of this TAM. And the HPC and AI workloads account for about 35% and the edge will ultimately be the largest at 45%. While these are rough estimates and there’s obviously some overlap, the bottom line is this market is enormous. And growing rapidly.

And you see that red highlighted area...that’s enterprise IT - traditional IT and that’s the x86 market in context. The prevailing approach of traditional IT vendors has been to re-package an x86 box and position it for the edge as a remote aggregation point. This is all well and good but: 1) It’s really not edge computing as many view it, it’s more like remote office installations; and 2) We believe the real opportunity is for AI inferencing at the edge where most of the processing is done locally.

General purpose x86 systems will play a role in this market, much in the same way that data center technologies factor in to mobile...but the real innovation, growth and action will occur much wider at the outer edge of the ecosystem in our view.

**Intel’s PC Dilemma**

We’ve made the point many times that PC volumes peaked in 2011 and we saw that as problematic for Intel for the cost reasons we’ve been highlighting. But amazingly, PC volumes actually grew last year thanks to COVID...and will continue to grow for a while. This is great news for Dell. But perhaps not as good for Intel.

Below is data from ETR that underscores that fact. The chart shows the Net Score or spending momentum breakdown for Dell’s laptop business. Green is accelerating spend and red is decelerating and the blue line is Net Score. The trend is up and to the right.
Now as we’ve said - this demand dynamic is awesome news for Dell, HP and all the laptop sellers. But it’s not necessarily great news for Intel. It’s okay but it shifts Intel’s product mix toward lower margin PC chips and squeezes Intel’s historically elevated gross margins. So Intel’s CFO has to explain that margin contraction to Wall Street. Imagine that irony...the business that got Intel to monopoly status is growing faster than the high margin server business and that it’s pulling margins down, which the street sees as a negative.

So as we’ve said, Intel is fighting a war on multiple fronts. Battling AMD in its core x86 business - both PC’s and servers, watching Arm mop up mobile, trying to figure out how to reinvent itself to allow more flexibility into its designs and spin up a foundry business to compete with TSMC.

It has to fund all this while at the same time propping up its stock with buybacks. Intel last summer announced it was accelerating its $10B stock buyback program. Hmmm. Buy stock back or build a foundry? Which do you think is more important for the future of Intel and the U.S. semiconductor industry?

Which is more important for the future of Intel and U.S. semiconductor competitiveness...Building a foundry or buying back stock?

Intel Must Protect its Past While Building the Future

And placate Wall Street at the same time.

Here's where it gets even more dicey. Intel must protect its high end x86 business at all costs. It is the cash cow and funds the company’s operations. Who is Intel’s biggest customer? Dell? HP? Facebook? Google? Amazon? Well let’s just say Amazon is a big customer – can we agree on that? And we know AWS’s biggest revenue generator is EC2 and EC2 is powered by microprocessors made from Intel...and others.

We found the following slide in the Arm Neoverse deck and it caught our attention.
The data comes from a data platform called Liftr Insights. The charts show the rapid growth of AWS’ Graviton chips, based on Arm of course. The blue is Graviton and the black, Vendor A, presumably is Intel, the gray is assumed to be AMD. The eye popper is the 2020 AWS EC2 instance deployments in the pie chart...nearly 50% Graviton.

Now if you’re Gelsinger – you are all over AWS. You do not want to lose this customer and you’re going to do everything in your power to keep them. But the trend is not your friend in this account.

**More Bad News - Graviton Speaks General Purpose Compute**

Below is the kill shot in the Arm Neoverse deck:
It shows the ISV ecosystem platforms that run on Graviton2. Because AWS has such good engineering and controls its stack, it can build Arm-based chips that run software designed to work on general purpose x86 systems. While it’s true the ISV has to do some work, large ISVs have an incentive because they want to ride the AWS Graviton wave. Certainly the user doesn’t know or care. But AWS cares because it’s driving cost and energy consumption down and performance up. Lower cost and higher performance. Sounds like something Amazon wants to consistently deliver to compete doesn’t it?

And the ISV portfolio that runs on Arm-based Graviton is going to continue to grow.

By the way, it’s not just Amazon. Alibaba, Oracle, Marvell, Tencent were all part of the Arm announcement, doing similar things but customized for their specific purpose. The list grows.

**Digging into Arm’s Technical Progress**

The last piece of the Arm architecture story we want to share is the progress they’re making and compare that to that of x86. This chart below shows how and where Arm is innovating.
Let’s start with the first line above under platform capabilities. The # of cores supported per die or system.

A die is what ends up as a chip on a small piece of silicon. Think of the die as the circuit diagram of the chip. These circuits are fabricated on wafers using photolithography. The wafer is then cut up into many pieces, each one having a chip. Each of these pieces is a chip and two chips make one system.

The key here is that Arm is quadrupling the number of cores. Instead of increasing thread counts it’s giving you cores. Cores are better than threads because threads are a shared resource. Cores are independent and much easier and less complex to virtualize. This is particularly important in situations where you want to be as efficient as possible, sharing massive resources – like the cloud!

And you can see on the right hand side of the chart under the orange, Arm is dramatically increasing the amount of capabilities compared to previous generations.

And one of the other highlights to us is that last line – the CCIX and CXL support. These refer to Arm’s memory sharing capabilities within and between processors. This allows CPUs, GPUs, NPUs, etc. to share resources very efficiently; especially compared to the way x86 works, where everything is currently controlled by the x86 processor.

CCIX and CXL support on the other hand will allow designers to program the system and share memory wherever they want within the system, directly, and not have to go through the overhead of a central processor, which owns the memory. So for example if there’s a CPU, GPU and NPU, the CPU can say to the GPU “give me your results at a specified location and signal me when you’re done.” When the GPU is finished calculating and sending the results, the GPU signals the operation is complete…versus having to ping the GPU constantly, which is overhead intensive.

Finally, “composability,” means the system isn’t fixed, rather, you can programmatically change the characteristics of the system on the fly. For example, if the NPU is idle you can allocate more resource to other parts of the system.

In fairness, Intel is headed in this direction too, but we think Arm is at least two years ahead.
Implications for Nvidia

This memory innovation is huge for Nvidia, which today relies on x86. A major problem for Nvidia has been coherent memory management because the utilization of its GPU is appallingly low and can’t be easily optimized. Last week Nvidia announced its intent to provide an AI capability for the data center center without x86 (i.e. using Arm-based processors).

So Nvidia, another big Intel customer, is also moving to Arm. And if it is successful acquiring Arm – which is still a longshot – this trend will only accelerate.

The bottom line: If Intel can’t move fast enough to stem the momentum of Arm, we believe Arm will capture 50% of the enterprise semiconductor spending by 2030.

If Intel can’t move fast enough to stem the momentum of Arm in the data center, we believe Arm will capture 50% of the enterprise semiconductor spending by 2030.

How Can Intel Maintain its Leadership?

This won’t be easy but we think there is a path. Remember in one of our earlier posts, we said Intel can’t go it alone? And we posited that the company would have to initiate a joint venture?

We had a triumvirate as shown below of Intel, IBM with its Power 10 memory architecture, and Samsung with its volume manufacturing expertise.

A Unique Joint Venture Could Emerge

Upon further review – we’re not sure that Samsung would be willing to contribute IP to this venture, given its huge investment in IP and infrastructure in South Korea. And furthermore we’re not convinced that Arvind Krishna, who we believe ultimately made the call to jettison IBM’s microelectronics business, wants to put his efforts back into manufacturing semis.
Breaking Analysis: Why Apple Could Hold the Key to Intel’s Future

So we have this conundrum. Intel is fighting AMD which is already at 7nm. Intel has fallen behind in process manufacturing, which is strategically important to the United States, its military and the nation’s competitiveness. Intel is behind the curve on cost and architecture and is losing key customers in the most important market segments. And it’s way behind on volume.

Intel must become more price and performance competitive with x86 to protect its core customer base and bring in new composable designs that maintain x86 compatibility. And provide customers the ability to add and customize GPUs, NPUs, accelerators, etc. All while launching a successful foundry business.

Whew.

So we think there’s another possibility to this thought exercise.

Apple Could Hold the Key to Intel’s Future

Apple is currently reliant on TSMC and is pushing them hard on 5nm. But Apple needs a trusted onshore supplier. Sure TSMC is adding manufacturing capacity in the U.S. in Arizona, but back to our precarious scenario in the South China Sea. Will the U.S. government and Apple sit back and hope for the best? Or will they hope for the best and plan for the worst? Let’s face it. If China gains control of TSMC it could block access to the latest and greatest process technology.

Apple just announced it’s investing billions of dollars in semiconductor technology across the U.S. The U.S. government is pressuring big tech. What about an Apple Intel JV? Apple brings the volume, its clout, money, design leadership to the table and partners with Intel. It gives the Intel foundry business a guaranteed volume stream and maybe the U.S. government gives Apple a little breathing room on the whole break up big tech narrative.

Wouldn’t that be ironic? Apple dumps Intel in favor of Arm for the M1 and then incubates and essentially saves Intel with a pipeline of foundry business.

A Unique Joint Venture Could Emerge

IBM in this scenario maybe just gets in the way so why not a nice clean partnership between Intel and Apple. Who knows, maybe Gelsinger can negotiate this without giving up any equity.
But Apple could be a key ingredient to a cocktail of a new strategy under Pat’s leadership, gobs of cash from the U.S. and EU governments and volume from Apple.

Still a longshot but one worth pursuing….because as we’ve written. Intel is too strategic to fail.

Keep in Touch

Remember these episodes are all available as podcasts wherever you listen.
Email david.vellante@siliconangle.com | DM @dvellante on Twitter | Comment on our LinkedIn posts.
Also, check out this ETR Tutorial we created, which explains the spending methodology in more detail.
Watch the full video analysis:

Image credit: jukovskyy

Note: ETR is a separate company from Wikibon/SiliconANGLE. If you would like to cite or republish any of the company’s data, or inquire about its services, please contact ETR at legal@etr.ai.
David Vellante is co-CEO of SiliconANGLE Media, as well as co-founder and Chief Analyst at The Wikibon Project, the world’s leading open source technology research community. Dave is a long-time tech industry analyst, entrepreneur, writer and speaker. As co-host of theCUBE - “The ESPN of Tech,” Vellante has interviewed over 5,000 experts since 2010. He is also a co-founder of CrowdChat, an angel funded startup based in Palo Alto using big data techniques to extract business value from social data. Prior to these exploits, Dave founded a CIO consultancy and spent a decade growing and managing IDC’s largest business unit. He lives in Massachusetts with his wife and four children where he is active in town activities including serving as the president of his town’s local “Kiddie Sports” association. Dave holds a B.S. in Applied Mathematics from Union College.