CTO-Vision-Panel at FMS

A vision of the Future: Insights from Leading CTO's

Thursday, August 8, 9:45 am - 10:50 am at the Santa Clara Convention Center

Panelists

Rita Wouhaybi (Solidigm) is an influential figure in AI and edge computing, currently driving AI strategies at Solidigm. With a background in Intel's CTO organization for network and edge, she has led groundbreaking projects in manufacturing, pharmaceuticals, and warehousing. Rita is passionate about democratizing AI and enhancing human capabilities through continuous learning and edge AI innovations, positioning her as a visionary leader in the field.

Al Fazio (Intel) Senior Fellow in the Foundry Technology Development Group, overseeing Intel's memory strategy. With over four decades of experience, Al has been instrumental in developing the industry's first commercial flash memory, multi-level-cell memory, and 3D XPoint storage class memory. His contributions significantly shaped the memory and storage landscape, with over 30 patents and numerous awards for his technical achievements.

Manoj Wadekar (Meta) is a leading figure in Meta's technology division, focusing on advanced computing solutions & AI integration. His expertise in system architecture and data management has driven significant innovations in Meta's infrastructure, enhancing the performance and scalability of their computing systems. Manoj's forward-thinking approach and deep understanding of AI's impact on technology make him a key contributor to the panel.

CJ Newburn (NVIDIA) Distinguished Engineer driving HPC strategy and technical IO roadmap in NVIDIA GPU Cloud, focused on pushing the envelope for storage and networking programming models at scale. CJ is a community builder with a passion for building an ecosystem that extends the core capabilities of hardware and software platforms from HPC into AI, data science, and visualization. He tinkers with and leverages NVIDIA and vendor products in a lab packed with scaled compute, storage and networking gear to apply and extend new tech.

Paul Borrill (Daedaelus) is a seasoned technology executive and entrepreneur with a distinguished multi-decade career. Paul was CTO of VERITAS Software during the 9/11 Emergency Data Recovery. Paul cofounded the IEEE Hot Interconnects Symposium and the SNIA. He has extensive experience in the design of fault tolerant & disaster resilient infrastructure design at NASA, Sun, Quantum, VERITAS & Apple. Paul's obsession is making transactions fast, reliable and secure.

Moderator

Russ Fellows (Futurum Group) with over 30 years of experience in the IT industry developing, operating, and managing IT applications and infrastructure.

Summary

Otter Recording

The panel, hosted by Russ Fellows, featured several high-profile CTOs and industry experts, including Al Fazio (Intel), CJ Newburn (Nvidia), Rita Wouhaybi (Solidigm), Manoj Wadekar (Meta), and Paul nll (Daedaelus). The discussion covered a wide range of topics related to AI, HPC, memory, storage, energy efficiency, and networking technologies.

1. AI vs. HPC Needs

The panel began by exploring the differences between AI and traditional HPC requirements. Al Fazio highlighted the similarities in terms of the need for high-performance memory closely coupled with compute. However, he emphasized the scale of AI as a key differentiator, particularly in terms of energy consumption for data movement. Manoj Wadekar added that AI infrastructures require significantly higher reliability and availability than general-purpose compute infrastructures, given the risk of job failures in AI workloads. Rita Wouhaybi introduced the importance of edge computing in AI, arguing that it requires new forms of compute that differ from traditional HPC.

2. Energy Efficiency

Energy efficiency was a major theme, with multiple panelists discussing the challenges and opportunities in reducing power consumption in AI systems. Paul Borrill suggested innovations like "defragmenting" data centers and optimizing power usage by moving virtual machines onto fewer physical machines. CJ Newburn discussed the potential for AI to manage data orchestration more efficiently. Rita Wouhaybi pointed out the inefficiencies created by the ease of spinning up compute resources, which can lead to wasteful power consumption.

3. Persistent Memory and Storage

The panelists discussed the evolving role of persistent memory in AI and storage systems. Al Fazio reflected on the challenges of introducing revolutionary memory technologies, emphasizing the need for a supportive infrastructure and the difficulties of integrating such technologies into existing systems. CJ Newburn and others discussed the importance of extending storage as an extension of memory, particularly as data needs exceed traditional memory capacities.

4. Networking Innovations

Networking technologies were another key topic, with discussions on whether current technologies like Ethernet and InfiniBand are sufficient for future AI workloads. Paul Borrill shared insights from Bob Metcalfe, co-inventor of Ethernet, and the potential for networking innovations to drive future AI advancements. The importance of reducing data movement costs was highlighted, with the panelists agreeing on the need for tighter integration between compute and data through innovations like SmartNICs.

5. Democratizing AI

Rita Wouhaybi emphasized the need for AI to become more efficient and accessible, moving beyond just training models to inferencing at scale. The panelists agreed that AI's widespread adoption would depend on its ability to deliver business returns while addressing power efficiency concerns, particularly in regions with energy constraints like Europe.

6. Future Challenges and Innovations

The panelists touched on several future challenges and areas for innovation, including the need for closer collaboration between hardware and software development, the potential for applying AI to optimize data center operations, and the importance of developing sustainable AI technologies. The discussion also explored the role of emerging technologies like CXL (Compute Express Link) in providing memory as a service, though the panel was cautious about its feasibility given current memory reliability requirements.

7. Audience Interaction

The session concluded with questions from the audience, covering topics like data efficiency, the role of memory in AI, and the need for new approaches to data management in AI-driven environments. Panelists reiterated the importance of innovation in networking, memory, and data management to address the growing demands of AI.

Transcript

Russ Fellows 0:35

Welcome everyone. My name is Russ Fellows, and I'm going to be hosting this esteemed panel. Paul Borrill put this together and kindly handed it off to me so that he could also participate. We ave a lot of very high level CTO people here, a lot of doctors on the stage. We're going to start on the far right with Al Fazio, who is a Senior Fellow in the foundry technology development group at Intel. Thank you. Al, I'll let you introduce yourself in a minute. But So you've been instrumental in developing industries, commercial flash technology, multi level, cell memory, 3D X-point and contributions include significantly shaping memory and storage landscape with over 30 patents. Anything you want to add to that?

Al Fazio 1:29

No, happy to be here. Okay,

Russ Fellows 1:32

Great. Next up we have CJ go in order. Here, there we go. CJ Newburn, so CJ is from Nvidia, Distinguished Engineer driving HPC strategy and technical IO roadmap at Nvidia for the CPU cloud focused on pushing the envelope for storage and networking programming models at scale. CJ is a community builder with a passion for building ecosystems that extends the core capabilities of hardware and software platforms from HPC into AI, data science and visualization. He tinkers with leveraging Nvidia and vendor products in the lab, pack with scale, compute, storage and networking here. So CJ, welcome anything you want to add to that Al?

CJ Newburn 2:16

Sounds Great, thank you.

Russ Fellows 2:17

Rita Wouhaybi. So Rita is an influential figure in computing, currently driving AI strategies at Solidigm with a background at Intel CTO organization for network and edge. She has led ground making projects in manufacturing, pharmaceuticals and warehousing. Rita is passionate about democratizing AI and enhancing human capabilities through continuous learning and edge AI innovations. So her position allows her to work as a visionary leader in this field. Rita, welcome,

Rita Wouhaybi 2:56

Yeah, thank you for having me.

Russ Fellows 2:58

All right, next to last we have, Manoj Wadekar, from Meta, is a leading figure in Meta's technology division, focusing on advanced computing and solutions in AI integrations, his expertise in system architecture and data management has driven significant innovation in Metas infrastructure, enhancing the performance and scalability of their computing systems. Manoj is forward thinking approach and deep understanding of AI's impact on technology make him a key contributor to this panel. Anything to add?

Manoj Wadekar 3:33

No glad to be here.

Russ Fellows 3:34

And last but not least, like you said, the organizer of this panel, Dr Paul Borrill also is a seasoned technology executive and entrepreneur with a distinguished multi decade career. Paul was CTO of VERITAS software during the 911 emergency data recovery, and co founded the IEEE hot interconnect symposium and SNIA. He has extensive experience in the design of fault tolerant and Disaster Resilient infrastructures designs at NASA, Sun Microsystems, Quantum, VERITAS and Apple. Paul's obsession is making transactions fast, reliable and secure.

Paul Borrill 4:15

Thank you. It's a pleasure and an honor to be here amongst all of these technical experts in the industry. Thank you.

Russ Fellows 4:22

All right, so with that introduction in place here, hopefully we have everybody up there. All right, so we're going to talk about a few different areas. And I think probably one of the best high level questions to start with is, how do the needs of AI market differ from those of traditional HPC market regarding memory, storage and interconnect technology? So maybe we'll start down at the end. Al, you want to start with that? Any thoughts on that question?

Al Fazio 4:54

Well, you know, I think in many regards, a lot of the AI stuff is kind of evolved from the traditional HPC supercomputer type of market, and a lot of the similar attributes of needing very high performance memory very closely coupled to the compute. And so in some respects, I think there's a similarities. So I think there's a fair amount of similarity in that regard.

I think the thing that's kind of different now is really the scale that's taking place. And in particular, you know, if you think about between memory and the compute, what really is interesting to me about this is that in a memory cell, it probably takes a couple of Pico or a fraction of a picojoule per bit to read or write, you know, once you get through the circuitry and things like that, you move it a few millimeters, and you spent more energy moving that bit, just a few millimeters, a handful, less than five millimeters. Than you did reading and writing with all the circuitry that's around that and so in AI right now, the real issue is it's all about memory bandwidth and close coupling with that compute.

But how do you get the capacity that you need at that bandwidth? So it's a capacity-density, a compute-density. And if you're doing it at scale, it's now you're burning a tremendous amount of power just moving that data around. And so it becomes a very interesting kind of multi dimensional problem. Many of the similar things that took place, I think, in traditional super compute and HPC, but just, I think it's on steroids now,

Russ Fellows 6:52

right? So Manoj, maybe you have a slightly different perspective, being more software focused, I'm assuming, working at Meta. Any thoughts on that same, the differences in ?

Manoj Wadekar 7:02

Not the differences, I think Al put it really well. I mean, definitely there is a strong roots from HPC and that has evolved the AI. Maybe I can add a little bit of a different twist to this, from the infrastructure perspective, that for hyperscalers, it's not how AI is different than HPC, but also how AI is different from our remaining general purpose compute infrastructure, because that's where how we build the data centers, how we build our software, how we manage that whole infrastructure is fundamentally different for the whole AI space, primarily because most of the general purpose compute infrastructure runs with an expectation of failures. You have millions and millions of small, stateless tasks that you can make sure that you know some of them just fail. We have the availability in the software infrastructure to allow for that, so the software hardware can be relatively acceptable to be failing. We can manage it, the software is aware of it, and software can deal with it.

The AI space, the software is not the same way you have a task that runs across lots of compute units, which are the GPUs, and any failures in that job, any part failure in the part of that job can lead to the whole job failing, which may be running for weeks and months. So the reliability expectations and the availability expectations for the AI infrastructure are dramatically different. So But having said that, of course, the memory challenge and all of the challenges remain same for the system composition, but to run the whole infrastructure, there is a fundamentally different expectation from the whole infrastructure, and also going up to the component level, the memories, the connectivities, the networks, the cables, there is a significantly increased awareness of how the failures happen and how to deal with it. So that, I would say is some specifically different ones. I would say.

Paul Borrill 8:47

I hear two answers to this, this question across the industry. One of them is that HPC and infrastructures for AI somehow going to merge together sometime in the future, and the other is that they're going to adapt differently and have different tiers of networks that have to be accessing them. The North South network will be relatively consistent with what we're doing today with Clos networks, but the there are other things that can replace them are like direct networks that you used in at the very top, where you can get direct access, perhaps through a MEMS device from one node to another without any of the intervening switching delays. So I see this evolution going in several directions. Very interestingly, I'm looking forward to seeing a lot more innovation in this area instead of stagnation on old architectures.

Manoj Wadekar 9:39

Right

Russ Fellows 9:41

So really, do you want to pick up on that? Or, I have a slightly different question, but if you want to,

Rita Wouhaybi 9:47

I want to augment or a little disagree with the previous commenters. I think these days, when we talk about AI and we stop there, it's almost like we're saying computer sci-

ence, the field is so large, and we are ... the answers assumed just training and data center. And I want to remind us that a lot of the AI innovation is starting to happen at the edge, and that is going to look very different than HPC. As a matter of fact, I keep telling a lot of my colleagues these days and the customers that we spent as a community the last several decades moving data to where compute is, and now we're realizing that data is even more valuable than compute.

So the edge really is bringing back all of that innovation in new ways, where things are going to be distributed, data is going to be distributed, and it's no longer where training happens once and inference happens forever, right? We're seeing a huge continuum of things between training and inference, whether it's continuous learning, whether it's tuning of LLMs. So really, comparing it to HPC is, I think, very restricted to only some of the training that's happening at large scale. But think about all the interesting things that are going to happen in factories and hospital rooms on the road with autonomous driving, and it looks nothing like HPC. It's actually a new way of compute that is so riper for innovation. Okay, great. CJ?

CJ Newburn 11:24

I think you could say that rather than AI replacing HPC, it's actually building on HPC. So HPC, and definition of Thomas Schultz is, is kind of a godfather of HPC, is that that's about scaling up and scaling out with rigor and discipline. But I think if you see some of the of what's changed and what's different in HPC, typically the way you scale is that you're using weak scaling. So you do the same thing on bigger and bigger machines in the relatively homogeneous way.

I think what we're seeing here is that we're going after strong scaling, which is why we need a radically different kind of interconnect, something like NVLink, to be able to have lots and lots of compute resources work on the same part of the problem and solve that as quickly as possible.

And that radically changes your architecture. That's why you need a GPU that's the size of a rack, and it also shifts to your point, working at the edge as an example, but even side, inside a data center, you're going to need different kinds of storage technologies for different purposes, where you're ingressing your data, you're pre processing your data, you're doing training and inference in different parts, and each of those demands a different kind of technology, including in storage, and hopefully we'll get to more talk about that differentiation.

Russ Fellows 12:45

Yeah, definitely has some questions on that, on a slightly different topic. So looking at energy efficiency. Paul Borrill, I'll start with you, what possibilities exist to improve energy efficiency in AI and storage systems.

Paul Borrill 12:59

So I talked to a lot of people and changed my opinion actually, just in the last couple of days, because there's a lot of discussion, especially with with some of our colleagues are in the in the audience here, about what the cost is. It's reaching a point where the cost of moving bits is a lot more expensive in energy than the cost of computing bits. That's a big deal. And so I'm thinking back, what innovations have we seen in the past that we could look at like, for example, we used to defragment disks, you know, and no, file systems don't do that anymore.

But what's happening now is, you know, could we think in terms of defragmenting the whole data center? Or even, you know when, when there was this VM VMware came out with this idea of VMotion, which allowed you to move a virtual machine and then continue to address it and have everything continue to work with a very fast failover, and that can only work at layer two. You can't do that at layer three, unfortunately, which is why it doesn't work for containers, but that brings into mind the idea of like, well, if you can condense a lot of virtual machines onto fewer machines so you can power down the machines you're not using, that's one way of reducing power dissipation, but another one might be a more generic form of let me call it power gerrymandering.

Russ Fellows 14:24

So CJ, any thoughts on power efficiency in general?

CJ Newburn 14:29

I think we have a lot of people who are coming and joining this fray and want to be able to either use the systems that we give them or to develop new applications, but they don't know how to deal with power. They don't really know how to deal with performance.

So essentially, introducing layers of abstraction beneath which you can do innovation, I think end up being really important. So one of the I totally agree with what you were saying, Paul and Al, about sort of the movement of the cost of moving bits, that is that kind of problem is beyond most, most users capacity.

So being able to we're introducing some things that take kind of a serverless approach, where, as an application trying to get some work done, you don't particularly care where the data is, how it's formatted, or how close it is, or what it's called, you're just operating on the data, and you can relegate to a runtime system, and people that are really good at tuning, performance, tuning, energy efficiency, can manage data orchestration in the background, and that there's a lot of room for sort of democratizing that, and having lots of people innovate And decoupling that from the people that are trying to specify what should be done.

And again, I think that being able to this is a tremendous growth area for being able to get smarter about how do we sort of pre partition the data and protect predict what's going to be most effective with that? And I think AI probably has a part to play in that that's yet to be discovered.

Russ Fellows 16:02

Okay, so using AI for AI? Yeah, I've heard, I've heard that a lot of ways for that to happen, sure it exists. So, Rita, any thoughts on energy efficiency and ...

Rita Wouhaybi 16:13

Yeah, actually, CJ brought up a good point about, you know, it's very easy for the developer these days, to spin up a task or a job, right? We made these abstractions, and it's, it's fairly, fairly common for a grad student or an intern to spin up some compute somewhere, and it's consuming power. As a matter of fact, how many of you heard a story or experienced it yourselves, where you spun up compute in the cloud and forgot it there, and it ran for days, and then you got slammed with it with a bill. Yeah, it's painful,

but, but imagine that your compute that you forgot about it has been running, has been consuming power. It's moving the bits, but it also these abstractions and abstracted layers that we have created for my 13 year old to be able to spin up a hugging face model and do whatever he wants, costs money and power and resources, and honestly, a lot of people, it's, it's kind of, I hate to use the word, but it's a little bit of sloppy use of the technology. So we have to figure out how to be more efficient. I think AI is maturing and has become the killer app that it also has to become more efficient in power and moving bits. You know, everything that everyone have talked about, I fully agree with. And it's about, you know, wake up call these days, when everybody's seeing the bill, whether it's the power bill or the compute bill, to figure out how to be more efficient in our use.

Russ Fellows 17:46

So the Wouhaybi household is pretty strict, I'm guessing. So the 13 year old has to start spending a hugging face. Holy crap. Manoj, any thoughts on energy efficiency? I'm sure with, you know, 10,000 GPUs, I'm sure that's of no concern to Meta.

Manoj Wadekar 18:04

Oh no, not at all easy. But I think I wanted to bring back to that point of the discussion, actually, the efficiency, and talk about Al, that brought the point of the cost of moving the bits, which is what is changing a lot. If you think of taking the 10,000 or 100,000 GPU clusters, the amount of power and amount of cooling that is required to hold cool the whole thing is the important aspect. And as we talked about the moving the bits, if you as you make the scale up cluster or a scale out cluster, the moving the bits is the most expensive.

This is why we see that actually the the innovation is happening is to keep on bringing things closer and closer together. This is where, in the whole systems coming on to the wafers or the all in the die and coming much, much closer with the chiplet kind of technology is going to become more and more prevalent, because you want to bring the bits in closer if you have memory requirement, the memory bandwidth requirements for GPUs and easily in the 10s of terabytes per second, and growing, which, if you start moving them out, the cost is very high from the power perspective.

So you need to start bringing them together. That requires a technology, that requires manufacturing technology to get more die to die functionality working at a higher bandwidth, your Shoreline matching, and the capacity and of functionality coming from different vendors at the different technology points. So I think the lot of innovation is going to come, continue to come in this area in coming years, that will drive the energy efficiency by bringing things tighter and tighter together, the amount of transistors we can put per square millimeter is not growing fast, but how many things we can combine together is where the innovations will continue to happen.

Russ Fellows 19:41

Okay, so Al, you want to jump in ?

Al Fazio 19:42

I think this is a really interesting multi dimensional problem. Think about it from just the physics perspective of you not wanting to move things over a large distance, because you'll just be spending the displacement energy and moving wires around. But then you have a software problem, because you know, as you bring those bits in to a smaller location, you have a finite amount of capacity that you can act on. And so how do you partition the problem down to a small enough compute? This is one of the reasons that you know computed memory, or processing memory, really has never taken off the the compute that I have there doesn't have the data, so I need to have it over there.

And so it's a hard problem. This is not one of those easy problems to solve, because it's how do you get things physically closer to the packing of that the bandwidth issues in there, along with, is it a large enough cluster that can do meaningful work on that cluster, which is a large software problem in there. And you know, to be honest, that you know, if you're working on something as like an embedded system, you'll work on trying to figure out every cycle. You'll know, those of us old enough, coded in machine language or assembly language, you look at every cycle, but there's only hands full of people that do that.

If you really want a developer community, you need layers of abstraction, and those layers of abstraction actually add layers of inefficiency to that moving of data around. So it's a hard problem, both and to solve technically. But then how do you scale it out to a large development community?

CJ Newburn 21:26

You find the right kind of parallelism to use on your accelerator.

Al Fazio

Exactly

Manoj Wadekar 21:29

If I may just I think Al brings up a very good point, and this is precisely why the whole problem for energy efficiency, for AI clusters is completely hardware-software co-design, there is no more independently developing hardware and let software work on it. It's really the overall optimization, the memory part of it actually, you will have the bandwidth optimized functionality staying closer to the processing units, but then you have capacity. We move little bit farther. But software needs to be very much aware of what goes where, if it has to optimize the product.

Russ Fellows 22:02

Right, so, on a related note, so talking about persistent memory technology and comparing that to traditional storage and even traditional memory, and what advancements might we expect to have in that? So maybe I'll start with you. Rita?

Rita Wouhaybi 22:16

Well, I have a position that we have to think outside the confines of the Von Neumann architecture. And I think bringing compute and data as close together as possible is the way for scaling AI, and both, you know, with a co-design with software power efficiency understanding, how do we deploy at large scale? But if you think about it, this revolution actually has been ongoing for a while. I would argue that it started with the SmartNIC. When the networking community said, I can no longer route things efficiently

by always going through a central point. I have to move some of the intelligence to the network. And I think the same is going to happen, where we have to move the intelligence very, very close to where the data is, because, by the way, we have made it so easy for people to collect a lot of data, and now all of us are swimming or sinking, depending how you look at it in data, and not all the data is useful, so we really have to start at every stage, looking at the data and being able to process it along the way and understand how to route it intelligently, and that has to become very distributed in order to scale.

Russ Fellows 23:36

Al, I think this is probably a question that's near and dear to your heart. Any thoughts on persistent memory technology and comparisons to storage?

Al Fazio 23:50

There's a lot of scars. We were talking a little bit of just in the hallway ahead of this. I find persistent memory is near and dear to me. Having done a 3d X point, I think the challenge in a lot of these things, as we were discussing, is, how do you introduce something very revolutionary in a system. We knew about persistent memory in the 1950s with core memory, and then we kind of forgotten it. So a lot of the software infrastructure doesn't understand it. A lot of the storage industry doesn't really understand it.

When we were trying to, first work on, how do you come up with a very, you know, memory, kind of, like storage? Well, the storage kind of software infrastructure believes that, hey, it's going to be microseconds, and I need multiple replications of that software. And geez, I actually have availability centers that have to be those 100 kilometers separated for disaster. And now I'm limited by the speed of light in there.

And so if there's a whole infrastructure that has to change in there, and and so it's a it's a little bit easy, actually, to come up with the solutions, but then how do you build out that whole infrastructure that has to revolve around it? Right?

So that was my lesson learned. It's really it's less about the individual technology than it is. How do you bring along all the pieces associated with that? Particularly, someone says, well, you know, I already have all my software built in this data center. I can't turn that right. And so I think those are the things we have to think about. Is for any one of these is, you know, what are the tipping points that allow a lot of these other technologies that have to come along in order to enable some of these fundamentals?

Russ Fellows 25:53

Yeah, the hardware is only as good as the software that enables its use, right? I mean, it's more than just a device driver. It's the APIs. I know SNIA did some work around that, partially worked on by Intel Andy Rudolph. I don't know if you know him, [worked with him very closely]. Andy Rudolph hired me into my first job, and I still remember the interview questions he gave me. But anyway, so he did a lot of work on it. But yeah, that can't happen overnight, and it's still ongoing, and probably more work to be done. So CJ, any thoughts on the hardware, the software, aspects of persistent liberty.

CJ Newburn 26:23

In one of my last talks. I think I remember giving before I left Intel eight years ago. I needed somebody to talk about PMem, so I jumped in and talked about that. And what struck me at the time, and still strikes me, is some of what you're getting at is, you know, as a technology, as technologists, we not only have to invent new technologies, but figure out who's going to use this and how much do they need to change how they're thinking.

And I think it's people aren't used to thinking of memory as being persistent. It's a little bit of a leap. However, I think there's a flip side, way of looking at that is, instead of changing people's thinking of using what's available, hey, persistent memory, disk storage as an extension of memory, and being able to say what I really care about as a user is data. And if I have huge quantities of data, then I want to ask, where do I put that? And one of the most effective place to put that from a total cost of ownership perspective, is to actually use NVMe.

So this is one of the things that we're seeing with this new classes of applications that need more data than can possibly fit in memory, is extending and spilling that over into this. There are times when you need to do that in a persistent way. You may need to draw in data that came from something else as a different phase, and you've used storage as an intermediate phase to transmit from one stage of a big pipeline to another. And persistent storage, has a great role to play in that. Or you may need to archive something, and those don't really fundamentally demand that people come up with a radically different way of thinking about data.

Russ Fellows 28:09

Paul, any thoughts on use of persistent memory?

Paul Borrill 28:12

As the speakers describe how they see things? I see a lot of agreement in how this is going to happen, but maybe I can describe it as: there's multiple waves happening in change right now. Maybe each of us have a chance to see one or two major paradigm shifts in our lifetime. But when all of the planets are aligned, you can see there's something big going to happen soon.

The three things I see are the miniaturization of packaging because of chiplets, and how the economics of that changes the story. I see the way that AI in particular is changing the access patterns and the fact that we need to access things as graphs, much more than, say linear memory or linear access to storage.

And the third being the innovations in networking, which were possible, and we see a lots of things. Even Ethernet is being reinvented these days. And, of course, we've learned a great deal, from InfiniBand and from NVLink. So I see these, see three separate waves. Is all happening at once in superposition. It's like there's a tidal wave about to happen, that where the innovation can really start to come in.

Manoj Wadekar 29:22

Yeah, I just want going to continue the thought that Paul started, I think the innovation is going to be really rapid demand, at least, is very rapid. I think the challenge that industry will need to go through is basically the difference that in the past, has been extremely open innovation. The technologies like Ethernet or TCP or memory technologies, anything has been very open because they need to

interoperate in the systems at different level. I think the demand for performance and demand for speed, for that performance is so high that we are there is a risk of everything getting more and more closer and tighter, and hence more proprietary. So I'm just looking at that as a potential warning for us to see how the innovation can go fast.

Russ Fellows 30:09

right? So this is my question, which is essentially that, you know, the current state of the art with transformer technology and a little bit diffusion technology and AI, is that as the problem size doubles, the mathematical complexity, you know, grows exponentially by a factor of two, right? So in order to increase this problem space that's addressable, do you put more of the impetus on hardware, you know, just scaling up and figuring out ways to grow the hardware exponentially, or do you think there should be more focus on software to make, you know, come up with technologies that are hopefully big O of n squared more efficient, right? So any thoughts on that, I'll start with you. CJ,

CJ Newburn 30:53

yeah, the it's really clear they're on the order of five different kinds of parallelism that you really need to exploit in order on a scaled system. So we're in the process of bringing up systems that are Grace-Blackwell systems, where you can get 72 GPUs in a cabinet, and then you can lash those together with InfiniBand or something to create something with hundreds or 10s of 1000s of GPUs.

And it's really clear that I'm working with a lot of the guys that are essentially carefully crafting, looking at, what are the different kinds of parallelisms available, and to what degree, and for each of sort of the instance of the parallelism, what are the communication patterns? How communication intensive is that, and how do I map that onto the underlying system that I have?

And that goes back to this earlier comment about being able to use strong scaling wherever you can, and maximizing the bandwidth that you can get, and minimizing the latency, and being able to sort of reduce .. you talked about the abstraction, reduce the overhead by just being able to do direct loads and storage to those so it's very much I love. I wanted to reach out and say, let's do co design together, because working both sides of those problems, of seeing what it is that we need to do in the hardware and the software is critical to the success.

And we as a company, you know, are supposedly, you might think of this as a hardware company, but we're very, very, much a software company. Of the only way that you can make use of all of that hardware is through a tremendous amount of innovation all the way up and down the stack to make that successful. So it's a big barrier to entry.

Rita Wouhaybi 32:37

So yeah, I want to add to this. So I I'm in agreement with what you said, but what I don't like is the fact that when you started, you said, hey, when double my problem, when I double my data, the complexity is going to go exponential. And I think we should stop. And we're seeing academia, by the way, innovate big time in this space. We should stop just feeding the data blindly. I have numerous examples of customers, especially in manufacturing that I worked with, where they put a camera on the factory floor, and then next thing you know, they have 30,000 images. They have few terabyte worth of data, and now they have to go and annotate them. And no control engineer is going to do that. So what we need to look at is look at these images and say, well, 80% of them are of partial products and really boring conveyor belt images that no one cares about. We have to identify those and dump them and not show them to the user, because we reduce the cognitive load. We reduce the training time. We have a model that converges much, much quicker, and we have an inference that runs much, much faster. So it also like we have to do the smart things along the entire pipeline. And I don't think just continuing, unless you're doing LLMs Right, continuing to just feed data is always the answer,

Russ Fellows 33:58

Tight? So apply a little human intelligence into the processes.?

Rita Wouhaybi 34:02

Well, no, there. I mean, it doesn't have to be human, right? Because AI today is doing things like selective annotation, where you figure out through AI, you cluster your images, you group them, you're like, these five are the most representative. Let me start with those instead of showing the 50,000 and have them be overwhelmed,

Russ Fellows 34:21

Right? Yeah, Manoj, the hardware?

Manoj Wadekar 34:23

So I really like the point Rita made. So just thinking about it, I wanted to, but my thought process was also on the continuation of what CJ was thinking the hardware, software, co-design also goes on to multiple axes, right? In the sense there is a limit to how many transistors I'm able to put every time I have a manufacturing process improved, maybe I'll get 15% or 23% or something around that every time I go, which basically means the remaining has to come from some software level optimizations.

You go from different level. Go from FP 32 to 16 to 8 to 4. So definitely, your performance is growing, but you're trading off something. You're trading of accuracy. You're trading of the trading of things that that are acceptable at the overall solution level. Networking perspective, the software needs to be aware of the way the networking is, as you just said.

And you know, you have a scale up network, you have a scale out network, but you got you going to decide which part of the parallelism is going to work on your scale up network, which part is going on a scale out network. So we are coming to a level where hardware will continue to grow and enable memory. We talked about what stays in tier 1, what stays on to tier 2. That is again, more of a hardware and software decision.

So I think the the gains that we are going to continue to see going forward are going to be something that hardware and software decision are going to do together, at least in the AI space that I'm focusing right now.

CJ Newburn 35:43

Can I quick share an illustrative anecdote for that? Just to make that point, one of the things that I mentioned that we're in the process of optimizing for this next generation system, a fun game that's being played is we have some folks that designed a simulator, a simulation environment,

DL sim. You can guess what the DL might stand for, and we're noticing, hey, there's a discrepancy between DL's and what we actually measure with the software stack on hardware.

And the nature of many of the discrepancies are actually the people that designed DL Sim made very optimistic and aggressive assumptions about software being able to do operated speed of light and do really the best. And so the problem of sort of approaching the software and the simulator together isn't, oh, the simulator is broken. Let's make the simulator more accurate. It's actually the software is broken and it isn't optimized enough. We're finding pain points that we're stumbling across that are hurting us. Let's go fix the software to be more like those idealistic assumptions. I thought that was fun.

Russ Fellows 36:46

So Paul, any thoughts on the breakdown between hardware and software optimizations and where the focus should be?

Paul Borrill 36:52

The conversation was wonderful. What it made me think of is that we need to shift from like statically configured linear processing to statically configured array processing, to perhaps dynamically configured graph processing.

Russ Fellows 37:11

Okay, so Al, any thoughts?

Al Fazio 37:14

I think you know, going back to your original question about the exponential growth and the complexity, [right] with, with the data, I think there's, there's two parts. And, I enjoyed the Meta paper recently, on Llama, which I think did a good job of illustrating the point you were making. CJ, of okay, you know what amount of power? Yeah. How do you do the parallelization for what type of communication banner, if I need I think that was a nicely written paper to kind of do that. I think there's going to be more of that, which I would put on the evolutionary side.

But one of the things, I think, from a basic academia, that I think is missing is you look at all these various things of algorithms, and everyone comes up with the next algorithm, and you really have a hard time to test okay, is this at a limit? Where is it? And I started thinking, Well, why isn't there like a Shannon limit associated with with machine learning, or AI, similar to how there is on communication channels?

Turns out I did a little bit of research. Actually. Shannon, a year after his famous paper, wrote something on the entropy of language. And so I really think there's almost something that people have to go back to in academia, which is more of those fundamentals, and is there a very different way to approach it from a fundamental perspective. I do agree, in the near term, people are going to be doing the different levels of parallelization, map that to networks, map that to communication patterns. But if you really want to solve the problem, I think we got to kind of deal with it a little bit differently in the basic academia.

Russ Fellows 38:59

So change gears a little bit and stop talking about the scaling of compute and memory.

Let's look at the networking a little bit more. So you've been involved with Ethernet, and so broadening that out a little bit, do you know generally going to ask, are the existing networking technologies, InfiniBand, Ethernet, maybe NVlink, sort of, you know, sufficient. Or do the can those continue as evolutions, or does there need to be a revolution in networking technologies?

Paul Borrill 39:31

So I spoke to Bob Metcalfe recently and and he gave me some fantastic anecdotes. He said, what we do is we find the best networking technology we can every year, and we call it Ethernet. It turns out, the other thing you told me was also a very funny he said that because of all the opposition to Ethernet originally, a lot a lot of it from big, large, you know, established companies, they came up with this idea that Ethernet doesn't work in theory. It only works in practice. So to Al's point about, we have to go back to Shannon to understand what we mean by the theory here.

Russ Fellows 40:15

right? So, CJ, any thoughts on the networking from your perspective?

CJ Newburn 40:18

So we are doing, as you might imagine, it was a big step for us to acquire another company for some change in $7 billion to go hook things together, and that's kind of exploded since then. And there's we've just recently been talking about a number of innovations, of trying to address some of the shortcomings of Ethernet with respect to congestion and control and being able to support multi tenancy.

There are more things coming with that, I think, that are yet to be revealed and that you'll be hearing about in a while. I think it's also one of the interesting developments that have been for example, that happened here. First, InfiniBand has been able to put compute in the network.

So we have something called sharp where particularly for the all reduce thing. And like nobody really seems to care very much about all reduce but altogether, we put something in there for it that can essentially do compute in the network, so that you can do a reduction in the switches, rather than having to move all the data to the endpoints. And this gets to your point about sort of, how do you make things more efficient in terms of moving data? By figuring out where does data have to move, and what are the communication patterns, and how can we accelerate that

Russ Fellows 41:44

Sounds like graph theory. Al. Any thoughts on other networking technologies and evolution versus revolution that needs to happen?

Al Fazio 41:51

I think there's some two fronts. I agree that distributed computing, through the network of how you're processing things, is going to have value that's in there. I also think there's interesting aspects. Again, talking a little bit earlier about, hey, is optics, right? You know, having a resurgence. And again, if you're trying to, you know, solve a problem over a football field size data center, just cabling, of wires and all the aspects of that are going to be a problem, and does allow you to do some sort of whether it's optical computing or optical switching, or how do you bring optics in to solve some of those problems? I think are going to put

new twist that's on there. But there's another part which is about the network, which is, soon as you get into the network, you're getting into software protocol, as soon as you're getting into software protocol, you're no longer treating anything like memory, right, because it's just too slow, right? And so

CJ Newburn

And not that reliable.

Al Fazio

Yeah. And so how do you think about that? How do you not put those layers on that? And so I think those are kind of, I don't have a quick answer for any of that, but I think those are the problems that need the focus.

Paul Borrill 43:18

Put it in the FPGA in the SmartNIC.

Russ Fellows 43:23

did I miss him a CJ or not

Manoj Wadekar 43:26

On the network side, I'm just thinking about actually, since there are networks, have separated out your scale-out front end, and Ethernet has been in all the places. Except, as you said, another memory network side still seems to be now separating it out on the especially on the scale up side. And I think Paul brought up a good point, that Ethernet ends up basically adopting anything that is required, but it doesn't have the memory semantic. And maybe CJ, perhaps you have thought actually looking at because both of you guys have both of those fabrics. You have Ethernet fabric, and also you have this memory fabric, which is unwilling. I don't know whether Ethernet is going to get there. There are efforts in that area. Whether the scale up can run on the Ethernet. This is an area I think we'll have to see, because the networks tend to, as Al was saying, the you have this whole reliability to handle. That's why, typically Ethernet has TCP and RDMA, and everything that we do question is, basically, can compute deal with that scale up, kind of expectation of closest access to the memory, or does it require, continue to require specialized fabrics, as we have today? So we'll have to see that, right?

Russ Fellows 44:35

So I'm going to combine a couple questions here. So in terms of adoption. You know, there's a lot of interest, a lot of money being thrown at AI right now by companies not wanting to be left behind, but that's not going to last forever, right? So companies are throwing money at it right now, not wanting to be left behind, but relatively short order, they're going to want to see some returns, right?

That's one issue is, how do we get from the land of theory and a little bit of practice in HPC environments to practical business outcomes for, you know, Fortune 2000 and combine that with, you know, better power efficiencies, which aren't as big of an issue in the US, but I know in Europe, any of you have traveled to Europe, or have European customers that you talk with, we've heard from some of our European clients that not only are their data centers not able to grow power, they're being told they have to reduce their power consumption. So faced with a reduction, possible reduction of power and the need to get AI into production and producing something, you know, with business returns. What needs to happen in the next couple of years to make those things happen? Who wants to ? I'll let you start that Al.

Al Fazio 45:50

Well, I think when we think about AI today, and fear of missing out, or anything like, most of the things going in AI right now are about training, but you know, for AI to really be pervasive, it's, it's not going to be in training. It's really going to be in inferencing, which is a, you know, because you think about my CEO uses this analogy. How many of you develop a weather model? Yes, how many of you use a weather model? And so it's really about the inferencing and things like that. If you look at the workload, it's a very, very different workload. You look at the arithmetic intensity of a training workload versus the arithmetic intensity of a inferencing workload, particularly with transformers, which are all in the decoder. So it's all just vector, matrix, multiply, arithmetic, intensity of 2. It's a different problem. You probably need to a different way to go solve that which is going to lead to a different you know, performance, energy efficiency, dollars going after that. And so I think it's, you know, the it's not like, you know, all of a sudden I'm going to throw away my my transistors and my packet, my memory. It's just, how are you using it? How are you configuring it? What are you how are you optimizing it for where those problems will lie, right? So that you can make money on it without spending a gazillion dollars on power.

Russ Fellows 47:31

Yeah? So notice any thoughts on helping to further democratize, although minutes don't agree.

Manoj Wadekar 47:39

I think Al said that is correct. I think we are in the early phase of AI. So for the if I look at from hyperscale infrastructure perspective, for the general purpose computing infrastructure that we have, we have had it for many years now. So we are used to defining the systems. We are measuring the system, measuring the applications, and make sure that the platforms are optimized to free, not only the cost optimized, but the power optimized, because power is where our most of the challenges tend to be. The AI is in early phase. If I just look at the spider chart of different applications and what components they use to what level, it's completely out of shape, because we don't know yet. They are changing all the time.

They're still in the early phase. So we continue to make sure that, you know, we try to get optimization. But today it is mostly still investment, and so you try to optimize to some level, saying that, hey, try to keep on bringing memories closer together that we actually, because it's per meter millimeter, is going to be much more manageable. So there are at hardware level optimizations today happening. Try to bring it inside the rack, as we talked about, give lots of GPUs in a single rack, try to optimize it there. But I think the long term solution is going to be actually making sure that the SKUs that are defined are optimally designed. Today, when we have very memory oriented workload, we tend to put tons of GPUs because that's where my memory is, and I may underutilize GPU so as we stay into future, we have to make sure that, you know, we will understand our use cases for training and inferences better, and we'll optimize it for power, right?

Russ Fellows 49:07

So Rita any thoughts on this, although it sounds a little bit like your thoughts on the edge as well. Yeah.

Rita Wouhaybi 49:10

So I when, before coming to solid time, I spent eight years at Intel, and I worked with a lot of what Intel called end users, so factories and hospitals and so on to bring AI. And almost every time I visited a one of those customers, we started with an ROI discussion, return on investment. And, you know, one of the customers actually went public about the engagement, so I can mention them. We did a big project with Audi manufacturing, and I went and visited them in Germany, and we spent three days walking through use cases. And there were so many use cases that as a data scientist, I was salivating. Oh, this sounds so cool, but we don't usually, in businesses, do something only because it sounds cool.

So you're right in the sense that today, we're doing a lot of things in AI training very large models, because it is cool and because we're trying to advance the technology. But at some point, all of those decisions are going to go back to ROI. We're going to find those use cases where it makes sense, but hopefully with democratizing using models and using a lot of innovation and awesome hardware, we'll get to a point where more and more use cases will be affordable, and we'll be able to solve some stuff that has been sitting for decades. People thinking, Oh, I'm never going to be able to solve this. So we have to get to that discussion. And power is one of the equations, power, hardware scaling, human acceptance and so on. And obviously business relevance,

Russ Fellows 50:40

Great. So Paul, your thoughts on this? Oh,

Paul Borrill 50:43

I have many thoughts, but the thought that comes to mind is when I did my transition from being a pure nerd as a distinguished engineer hired by Sun Microsystems, and they gave me this job of managing the new product introduction process, and they said they needed more. Needed to switch to birth control instead of euthanasia. For projects, I took that on as a challenge, and I got some professional help, and I you'll hear me on the web being quoted. There's only one thing worse than an engineer who's never been to see a customer, and that's an engineer who's been to see just one.

Russ Fellows 51:25

That's a good one. CJ?

CJ Newburn 51:30

I think gone are the days when we had time to adjust to new tech. Things are come onto this treadmill area. There's a self fulfilling prophecy where every two years, the software models and the usage models and so on are just radically changing, and it's unsettling. And I don't know how many of us are having AI, write our presentations for us, or do our engineering work for us, but many of us are lagging behind that. But I think that there's, there's some adjustment required. You know, in the recent discussion Jensen had with Martin Zuckerberg, he talked about, everybody can have a couple personal models that can sort of figure stuff out for them or do whatever it is that they need to do for them. And that's a pretty new world.

And that's, you know, going back to the enabling discussion we were having, I don't know what that enabling is, because it's really reaching the whole world and our personal habits. So I think that that'll take us a while to catch up to. But doing the, you know, our investment is basically in doing a vertical integration of trying to understand some of the particular problems that people are working on, and one of them, Rita, will be talking about it at supercomputing, is some work we're doing with sort of AI at the edge of integrating because data, so much data, is coming in so fast. How do you figure out what's interesting about it? And how do you note, back to your conveyor belt example, how do you have a model of what you would expect so that you could notice something that's different, right? I heard a story about somebody who came from Japan to Oak Ridge National Labs, and they set up this experiment looking at microscopic features in silicon, and looking at some processes under temperature and pressure. And they got all the results, and they went away and analyzed them and said, you know, but, like, only 10% of my samples were interesting, because that was an anomaly, and if I'd only had a model that could have recognized that, and better spent the time that I had there on that instrument to be able to do it, I'd be able to come up with, you know, much faster results, because now we have to wait six more months to get my turn on that machine and that that infrastructure of being able to discern between what's normal, where are there opportunities for innovation, and what's something that I should take notice of, I expect to become much more pervasive for us, but we'll see what happens with that. It's difficult,

Russ Fellows 54:07

All right, so now time for our lightning round, the three questions. So the three wishes. So Paul has let me become a genie, so I get to grant each of you three wishes, and Genie rules apply. You can't ask for more wishes, and you also can't wish something for your existing products that you've designed. So you know, thinking outside the box, not limited to your role or your company. In 90 seconds each, we'll start with you. Al, what are your three wishes? If you had anything that you could wave a magic wand for

Al Fazio 54:42

If I had anything to wave a magic wand, and it's not to bring me a bucket full of cash. The first thing is, I would like every hardware or process technology engineer to understand a lot more about software. And I'd like every software engineer to understand a lot more about the hardware and processes. I wish we had a lot more cross disciplinary capabilities that used to be, I think, a lot more in the industry. And as we became specialized, they started to go separate. And I think it's time to bring that back. So that's actually my, my first wishes in the engineering community. The second wish I would have, is, is there a way to for this whole question about legacy and how to fit into that, which is really drives, you know, the need for evolutions as opposed to revolutions. You're not going to make a leap forward without that. And so some way to go solve that, and I don't know, okay, know what that that answer is, because there's an economic reality.

Russ Fellows 56:01

So one way so that we've solved it, it's your last one.

Al Fazio 56:06

The last one is probably, go back to what I was saying earlier, is that I think there's a lot of work in various algorithms today, but no one, I think, understands where the fundamental is of just, you know, what is an optimized algorithm? Am I at the equivalent of a Shannon limit? And if you understood that, then I think there could be, you know, more fundamentals about how you approach the problem. You would understand the workload, and therefore how to optimize the balance in the system between compute, memory and communications, better than we're probably doing today.

Russ Fellows 56:45

So we all need multiple PhDs. He's saying, okay, so Paul, what are your three wishes?

Paul Borrill 56:52

I came up with this question, and then I forgot about just now, so I'm feeling what I'm hearing here. And I think, number one, I'd like to see a lot more co-design between software and hardware, and more to it than that. I think from the APIs to the bits on the wire, we need to have much better and more disciplined, mathematically, formally provable, perhaps, relationships. So APIs to bits on the wire, that's that's the Shannon thing, basically. And another aspect of thinking about that is, instead of just having Shannon as a one way channel, which it typically is described as, it's really a bi directional synchronization, and we can do that on our modern networks. That was the second thing. The third thing, I think, would be, I'd love to solve the power dissipation problem. It's dear to my heart. I was one on the one of the early experiments that measured the sea surface temperature at University College London. I actually did some of the testing when I was an apprentice there, designing satellites and stuff to go in space. And I became sensitized to global warming a long, long time ago, and it's actually the thing that I am concerned about the most, and I do think as an opportunity. As we switch from the way we do a linear and an array kind of processing to graph processing, we actually can use applied graph theory to move things around and to have relationships that not only make the computation a lot faster and lower latency, but it also can do this. Let me say power or energy gerrymandering.

Russ Fellows 58:29

All right, great, Manoj, what are your three wishes?

Manoj Wadekar 58:33

Sure, I think I'll start getting more tactical and maybe specific about what we look for, at least in next, say, five years. I'm not going to go beyond that, maybe five to seven years. One thing which I see that, you know, I'm still seeing stick the Al's point of Pico joules per millimeter, still sticking with me. I think we need to continue to see, we would love to see where the technology allows us to be much more efficient, especially this is will be in the area of more and more packaging, together with the Chiplets, etc. Because I think that this is where our Moores law is leading. We are going to have lot more things working on a system on a wafer, kind of a so more technologies, more open technologies, collaboration, which basically brings me to the point of the openness aspect. But before that, actually, just want to talk about that. The second part is basically as we start going to start seeing this larger and larger components working together. It also comes to Paul's Point, but also power delivery. Power delivery is going to be challenged. We're talking about racks, basically having GPU the whole rack, basically how much power to deliver, how much each component can be delivered, and, of course, how they can be cooled. So the cool technology, important technologies, are going to be in this space. So we look forward to that solutions

For all of this. I think it's going to be critical to have openness in the ecosystem getting better. I think we are getting more and more proprietary as we go into look into the system. So I don't know how, but definitely sharing up more data about use cases, which comes from people like us, but also for implementation choices, technologies that allow multiple chiplets and everybody to work together, networking technologies to work together. From the problem perspective, I think these are the three things that is going to be important for us to deliver most efficient systems going five to seven years from now.

Russ Fellows 1:00:10

Okay, great. CJ,

CJ Newburn 1:00:14

Think out of the box. So one of the things that I have really found compelling for me, is to think with more of an abundance mentality, rather than a scarcity mentality, and of looking rather than how can we sort of protect ourselves, self, assert ourselves, and sort of grab with what's best for ourselves, which tends to divide us and separate us? Of looking at, where can we build connections, where can we bring the most out of other people and have a greater sense of optimism?

The second thing is an application of that. I think that there's an opportunity for us in this storage space, for us to look at what the new requirements are for applications as that's being reflected in what we need for technology, including with storage. Had lots of discussions with storage vendors this week of looking at, for example, how we can get more fine grained IOPS out of the media and the storage controllers. So specifically looking at how can we together, discern set up a framework where across a number of vendors, and for the sake of the industry, can discern what do we need and how do we best fulfill that, and where maybe we need to push standards forward in that space.

Just another very different thing that is a concern for me in this country and as a citizen here, if any of us thinks that any of our data is protected and secure, we have another thing coming. And the only way forward, I think, that we have for that is for us as technologists to work towards being able to achieve security by default. And that means that it needs to be low overhead. It needs to be the easiest thing to do, instead of requiring additional extra special effort to make that happen. And I think that there's a lot of movement in that space for being able to advance that, but we have a long ways to go.

Russ Fellows 1:02:12

So I guess CJ said the NSA visit, okay, it's probably most of you had him at one point, and Rita, I saved you for last.

Rita Wouhaybi 1:02:21

Thank you. So in typical engineering fashion, I have a clarification about the requirements

Russ Fellows 1:02:32

I've been on both sides. Product Marketing Manager,

Rita Wouhaybi 1:02:36

you never said they have to be realistic right? Well, I mean, for purely selfish reasons, I would like to see AGI in my lifetime, generalized intelligence. I think that would be cool. Yes, it is going to change the way we work and the way we think and the way we learn, but gosh darn it, it would be so cool to see it actually happen.

The second one is kind of, you know, have been mentioned few times, and that is the concept of sustainability, right? We are making this globe warmer. We as a community, are heavily contributing into that. We're consuming power. None of it is sustainable. None of it is recyclable. It is getting very, very scary.

And the third one, I think the AI community and us as memory and storage, we are so focused on saving every bit of data, but at the same time, us as humans, one of our superpowers is the fact that we can forget things. And I don't, I don't think, especially with the FOMO right now, that we're even daring to poke at that, you know, big bubble. So it would be good to start to understand what it means to forget some data, because it is, in the long term, more sustainable. You know, that's how our brains work. Yeah, and I think those are my three.

Russ Fellows 1:03:56

All right, great. So I think we can open it up to questions now, all right, you were ready?

Audience Participant 1:04:02

Yes. My name is David Shader?. I'd like to commend the gentleman from Nvidia who said, If you think your data is safe, you better think again. The forthcoming quantum candidates coming out of NIST are not proven to be unbreakable, therefore they are not secure by design. According to Jen Easterly, we are all ladies and gentlemen. Crash dummies. Booth 646. Combines quantum qubit superpositioning With listen for it. Shannon perfect secrecy, which has been mentioned many times. If you want to improve the ROI of your products, help your customers, help your customers protect their data with a data encryption technology that is not breakable by any method that is known other than brute force. And I'll be happy to talk to anybody, and I'm happy to come to Nvidia. I'll be happy to come to Intel Solidigm anywhere you can all throw rocks at me, I'll be the only person left standing. If there's any questions, fire them.

Jean Bozman (from the Audience) 1:05:13

I have a question. It just strikes me that it's one of those wordy things. What did you know? And when did you know it? And why you say that is, it's all your comments are. What do we want out of the system? What's our perspective? And it may not be a one size fits all thing, right? I mean, there are sweet spots here, here, here, here. How do we choose among them? That's kind of what I want to hear from you guys, because we can't do all of this all the time. What are the sweet spots? Any comments there where the wordly thing is just too much?

Manoj Wadekar 1:05:54

Wow, I can take first crack at it and then hand it off. It's very difficult for AI space right now, I think. But of course, there are rule of thumb saying that you know how much compute and how much memory and how much network, etc. So you make a decision from from now to next, what I want, but is it the most optimal something? I think that data is still kind of unbaked on that. For the remaining infrastructure, we have lot more historic data, and we understand, at least for the private clouds, we understand those use cases a little more. But AI think is still evolving from the optimization side.

Al Fazio 1:06:26

I think that the fundamental challenge in that question is that if you know what that is, if you just write that down and say, Okay, here's the design point, I want go off and do that. Okay, I'll get a design team go build the ASIC assuming that the technology's there. I don't have to invent anything in the process, technology or packaging. It's still a couple years out before I have something. Is that still going to be what I want a couple of years from now? That's generally the rule of why general purpose computing tended to be more pervasive than than single point solutions.

Now that doesn't mean that's always true. You know, in the case of GPUs, it was specialized, you know, applications. But, you know, over 20 years became, you know, a broader class that's in there. And so I think it's you have to be careful on the answer to that, because the more bespoke that you get, you're probably going to be wrong when it finally comes out, where, where the answer will be. And so it's, how do you bridge between something that's large enough and general purpose enough, but then you just it's not completely inefficient at that application.

Jean Bozman 1:07:41

Okay. Thank you very much.

Bill Gervasi 1:07:47

Bill Gervasi from Wolley, and if I had a wish, it would be that data would know if it's actually going to be used, then if it's not going to be used, it would just sit there and refuse to be moved and, you know, comes to Rita's point. Al hit on this, Paul definitely is talking about... how do we improve the efficiency of the data centers? I spent the last few years working on a report for the department of energy that is going to the Congress next month on the efficiency of the data centers being so close to zero that we should all be embarrassed. So how do we get there from here? How do we get data to stay where it is if it's not going to be used? We have these cache lines, we have block transfers and so forth, and every single level has a phenomenal amount of inefficiency. DRAM cache fill as just a single access to a row of DRAM is 0.025%. Average use of 100 bytes of a 4KB block is only using 3% of the block transfers. Where are we going to put together the technologies to tackle this problem of raising the amount of work we do per watt, and the results per watt, as opposed to gigabytes per second?

Russ Fellows 1:09:11

Let me add onto that briefly, because, you know, I was in the computer storage industry for many, many years, and

seen a lot of numbers relating to that. And you know, they vary somewhat, but not greatly, for decades. And that's like around 75-80% of data is WORN, Write Once Read Never. So there's a fast quantity, you know, whether it's maybe it's only 60% okay, still, whatever. If that number is anything more than 10% it's horrible, right? So, yeah, what do we do about that problem?

Al Fazio 1:09:41

I think this is a case where, you know, AI as a tool of embedding in inside of whether it's circuits or in systems, would be useful. You know, today, you know, any computing system anyone builds is going to have caching policies, right? What do you evict? What do you keep? What's warm data, what's cold data? And you can kind of, it's not too much of a stretch of imagination to say, hey, if I had more intelligence built into that, those sorts of policies that you can start saying, you know, the probability of that data being used, should I be pushing that out to a cold storage should I be pushing it out to a persistent layer? Or is this, you know, Write Once Never Used, great. I'm not going to ever access it, right?

Russ Fellows 1:10:34

Well, yeah, that's a good point, because. I've been in IT admin forever as well. Started in the 80s, and I'm still doing it today. And a lot of times you don't delete data just because of fear of management coming back and saying, Hey, what happened to that? I needed that. So because it's almost almost free, but not free and energy and efficient to keep it, most people, by default, just tend to keep things for fear of deleting something. But you're right, if we had some intelligence saying we really don't need that. And you can put your trust in it and say, Okay, I'll let AII decide to throw it away.

CJ Newburn 1:11:07

Then, until you run out of storage and your mailbox is full and you have to wholesale dump all kinds of stuff, maybe there's an opportunity riffing off we what you've said Al both to be able to, like if we do collect data, to throw it away as soon as possible, and to dispense with it, or at least make it extremely cold. But maybe there are also opportunities, again, to use your video example of maybe we detect at lower resolution, and when we see something that's potentially interesting, that we go back and increase the resolution and pay more attention to those things and then do before we do the post processing, so that we can filter some of that out at the front end. Maybe there are opportunities there.

Rita Wouhaybi 1:11:47

Yeah, I think adding intelligence as close to the data as possible, actually is the answer, right? And I have plenty of examples when someone would collect data in a factory, because maybe they're looking for defects, but then they realize that this camera is pointing in the background, and I can watch the workers who aren't wearing their hard hats for their safety and alert them. But if they had dumped that data, they can't go back and look at it and train with it. So adding intelligence to the data, where it sees what data is diverse and is able to increase and decrease the resolution on it in a smart way is essential. And I think treating the data as dumb and assuming that we're going to process it later is really an assumption that we're not going to be able to continue to live with.

Russ Fellows 1:12:35

Okay, maybe we have time for one more question.

Pankaj Mehra (audience) 1:12:39

I'll keep it brief. Then we enter a distance, and what's happening with CXL. What's your general thought? Where is memory useful as a service? It's the last infrastructure piece that we don't have as a service. Does it make sense to think about memory as a service, or will it always go with Compute?

CJ Newburn 1:13:02

We be controversial here, perhaps, but I think memory, by its semantic can't fail like you go and load something, you've got to get an answer back. And if you your whole contract that you made with the system is that you're going to get an answer. So if as soon as things get far away enough to be unreliable, then you need a lot of buffering to be able to make that work. And so there's a real I think we have a bunch of challenges ahead of us for both being able to manage the buffering, which is inefficient and involves moving more bits and buffering more bits and storing more bits, and managing that reliability for being able to make that work. So that may be a theory. That's one from practice. We'll see how that turns out.

Manoj Wadekar 1:13:45

Yeah, no, I agree with you. Maybe then the concept of what is that memory can change. If there was a way for me to say that, hey, I want that something, but that doesn't help with all raw data, but specific information or insight. I'm looking for that, but that requires a significant changes, of course, because it changes on the way what you ask the question, give me that frame versus saying that, give me the frame with some specific information that will move the intelligence towards that memory. But then your interface changed to it, saying that your question, the way you're asking, is changing. So depends on the definition of memory. I think then the memory distance will become interesting. Yes,

Audience participant 3 1:14:28

I have one question real quick, okay, is this is a great panel. You know, it'd be great to just sit down with you guys for dinner and go over this, because it was very interesting. But one of the thoughts that just keeps coming back in my mind is we've created AI today through LLMs as like a hammer, and everything looks like a nail.

And in listening to Rita, a lot of I think what we're missing here is where, when you look at LLMs, then you have this huge explosion of Ferraris, like engines and everything like that. But a lot of the work that gets done is going to be through supervised learning, right? Or we're going to, we're going to take the foundation model, the LLM, and shrink it down for for specific applications. And that, I think, is going to change the way we look at AI and also data use. And, you know, we're kind of creating a large scale infrastructure that may not necessarily be necessary.

Rita Wouhaybi 1:15:34

So, there is somewhere in between LLM and supervised learning, and I think that concept is becoming more and more prevalent. It's the idea of AI that continuously changes and learns with you. So perhaps starts, you know, assisting a human, and starts by being almost stupid, and learns as time goes on to become a good assistant that you can rely on. And it doesn't it like I said, it sits somewhere between supervised, and LLM think about semi supervised and unsupervised and watching data for patterns that are emerging, but you're spot on. I think we have a huge spectrum, and most people are very excited about LLMs, because obviously they are really exciting. And then supervise is lot of work in many situations, because you might not have the data. Or, you know, CJ gave a good example, someone who found only 10% of the data. That's interesting. So think about for supervised, if that person had to actually annotate the entire data, they'll be pissed. They'll be really mad. So I think there is going to be a huge spectrum that is going to fill the gaps and address. Many, many use cases.

Russ Fellows 1:16:48

All right, I think we're out of time. But thank you, everybody. Thank you.