A Tipping Point:
The Next Revolution for Compute, Storage, & Networking

Stephen J. Sicola
Founder & CEO StorTrek Consulting

Tipping **Point**?

- Since the 1970s, the relationship between data and workload characteristics and infrastructure elements (servers, networks and storage kit) have been steadily changing -- but maintaining **balance**…
- This began to change with the New Millennium: Data volumes grew significantly as a function of the **"Data Democracy"** and **the Internet and world wide web**, server processors stopped doubling their cycling rates with each generation, networks could not keep pace with traffic volumes…
- By 2020, **the tipping point** will be reached
  - Scaling servers will become problematic
  - Data will be in the tens of zettabytes range courtesy of Multi-National commerce, IoT and Big Data Analytics/AI, and spread across a hybrid infrastructure of on premises data centers, private hosting services, private and public clouds, networks will struggle to maintain connectivity between locations
  - Some call it **"the Zettabyte Apocalypse"**

Time to **worry**?

- One would think so…
- Lack of attention to infrastructure requirements has put us over a barrel
  - Hypervisor vendors insist that hardware is trash, the "hardware abstraction layer" is what counts, software-defined everything…
  - Multi-core processors hampered by single-core architectures, I/O log jams at the RAW I/O end of the bus that is not rectified by faster storage…
  - Scale-out architectures providing very mixed results in terms of latency, throughput, and many other engineering metrics…
  - Cadre of smart hardware admins and designers is shrinking fast…

The "Hybrid Cloud" changes **nothing**…

- Pushes certain problems out to service providers, but adds cost and complexity without addressing fundamental problems that require historical understanding to resolve…
- The triumph of marketecture over architecture….
- Many companies can't or won't put any of data in the cloud, but the Cloud is is here none the less

Problems at **Scale**

- Today's world of Compute & Storage is Mess….
  - Server Sprawl with low utilization (even with virtualization!)….costs in all dimensions
  - Network sprawl for distributed computing with expensive traditional or modified traditional networks
    - Oversubscription of networks (North-South Traffic) vs. the need for more East-West traffic for distributed computing causing costs to skyrocket for effective performance for scale out solutions that so many need based on how servers are used today.
  - Software costs through the roof with costs per core
  - Storage life cycles different than server life cycles bring back the notion now of Dis-aggregated Compute and Storage (so the SAN is back?)
  - The Panacea of SSDs to assist in application performance
  - The ever growing need to 'archive' data and its explosive growth
  - The desire for more and more computing infrastructure for Big Data Analytics, Rendering, etc..
  - The Panacea of 'In-Memory Compute'

Problems at scale with status quo

- The Industry today seems sold on upgrading the speed of individual devices such as NVMe SSDs, NVMeOF, RDMA, and 'In Memory Computing' are the way to go.

- The same goes for Virtual Machines from the original to now Containers & Pods
- NVMe SSDs are very fast, but they have to go over much slower interfaces such as PCI (vs. CPU Memory Channels….just evolution vs a real change from using peripheral storage for random I/O volumes.
- When used in Distributed Computing, it works better than HDD, but there are more key elements tied between server and peripheral, like networks.

- NVMeOF is a good idea to eventually get more bandwidth, but today's plans for NVMeOF is to basically extend access to more drives over PCI, then over the network, especially as JBOFs vs Arrays

Problems at scale with status quo

- NVMeOF typically is tight coupling with all data volumes on peripheral storage…..this tight coupling causes much ore attention to detail as all vital data volumes on external storage especially when things break, let alone the cross-sectional bandwidth of the Network
- RDMA is a good idea, as maybe RDDA may be… but with today's methods of distributed computing, even though it has speed depending on what speed of network used as well as how many ethernet ports are used per servers. Latency adds up!
- The issues with NVMe and NVMeOF, let alone any peripheral storage interconnect means that when memory is moved between potentially multiple servers, each of those servers have to store memory on to the peripheral storage devices, still with a tight coupling with storage.
- Even 'In-Memory' Computing involves peripheral storage and is not talked about much, because of the same paradigm, just with more RAM, meaning that much more access to peripheral storage and all the problems noted above

Problems at **Scale** and Public Clouds

- Problems at scale are hitting the Public cloud in a **big way**!
- The original Public Clouds (Google, Amazon, Facebook, Alibaba, etc..) were based on a **software idea first**, with software people putting together data centers.
  - Google…search engine based on work at DEC/Digital Research Labs (Alta-Vista), Stanford, etc……now a business cloud
  - Amazon…Web Shopping software….now a business cloud
  - Facebook…a Web people to people information sharing software….some business work here but the technology used (including OCP) is not performance oriented per se
  - Alibaba…a copy of Amazon, but now offering business services…..and many more like IBM!!

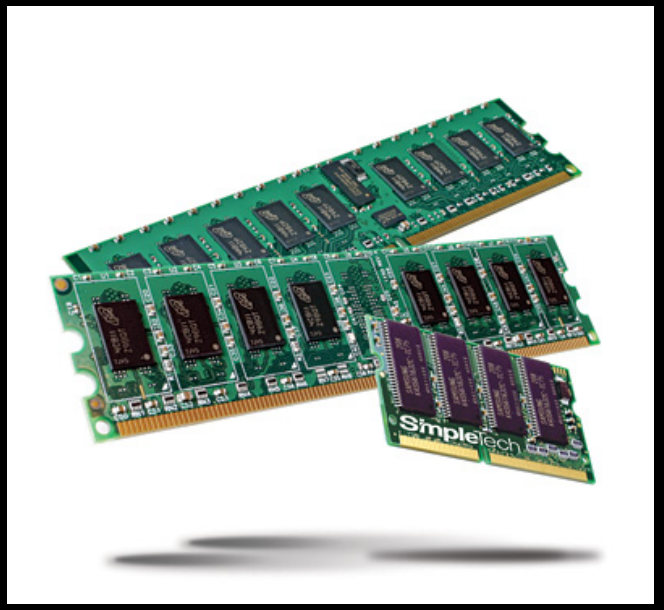
The Public Cloud is here to stay, though it's a proprietary design!

Public/Private Clouds

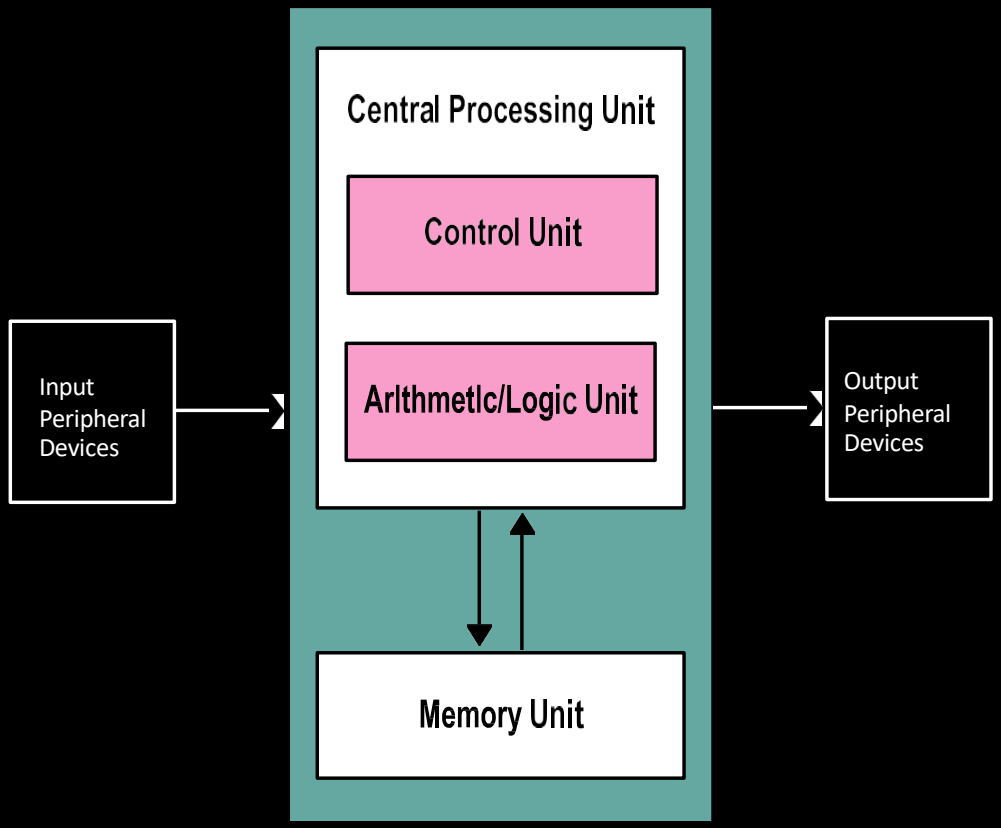
- Other Public Clouds that came **after the initial software-based** web clouds include Azure, IBM, etc. and many others across the world now.
- **Private** Clouds are now possible and hitting the same issue of sprawl and costly implementation with evolution vs. revolution for both types of clouds
  - Private Clouds now possible in many cases as older legacy servers running legacy software have been aided heavily with software that orchestrates their usage, hence a private cloud.
  - Other Data Centers still not at the private cloud stage **will** get there, but will they face the same challenges that drive IT costs/Public Cloud usage costs up because of the overall TCO.
  - ITS out of control…..this is called Entropy…If you think about it, Facebook/twitter, etc.. have reversed entropy by allowing humans to connect more with less…so why can't we have this for Data Centers, Cloud-Based, Hybrid, or Private?

A look to the past for an answer?

- Since the first DRAM (Volatile…loses contents when power goes away) was produced in 1970, with most computers using DRAM by the late 70's, it changed the way we compute and the way we use storage devices completely. Memory is just a cache and must be written to persistent peripheral storage (HDD, SSD, new memory-like DIMMs, e.g. Intel Optane as well as others, e.g. MRAM etc..).
- Basically, a computer was now split into 'server' and storage (Direct attached or externally attached Storage or Storage 'Arrays.'
- Software changed forever more to use Memory as a cache and use of HDD and now SSD have been part of application operation fully ever since.
- A look back at the original computer architecture is in order:

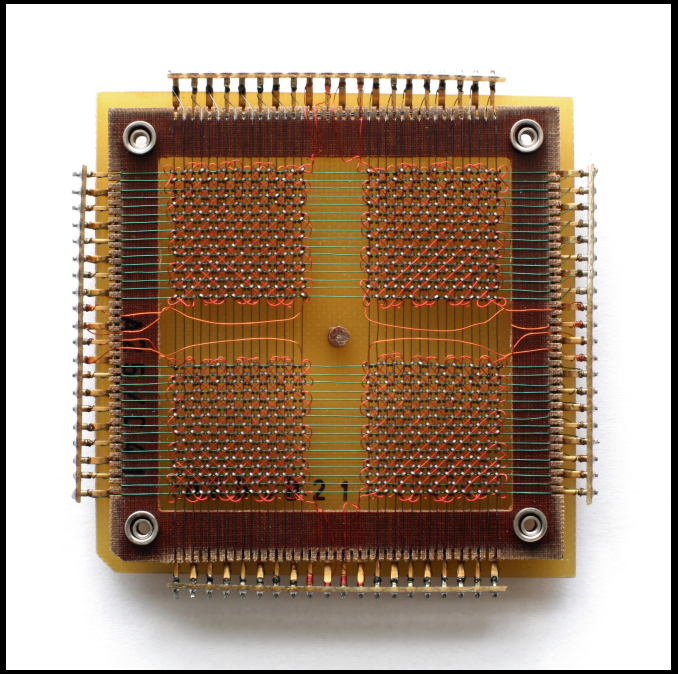


The original Computer Architecture (Von Neumann)



- Memory was memory….that means it was NON-Volatile and never went away
- All of Memory was used for applications as well as to hold Data volumes for applications to run against
- Input devices were punch cards, then paper tape, but also tape drives….e.g. figure on next page of early main frame
- Later rotating Drum Drives for input data and output information, and then the first rotating discs up to today's 2.5 and 3.5", older SSD's made of faulty DRAM, ECC chips, two HDD, and a DELCO battery (TRUE), to SSD's of today.
- Mainframes and Mini-Computers as you see has NVRAM based systems and adhered to the pure Von Neuman Architectural model

Mainframes with NV Memory

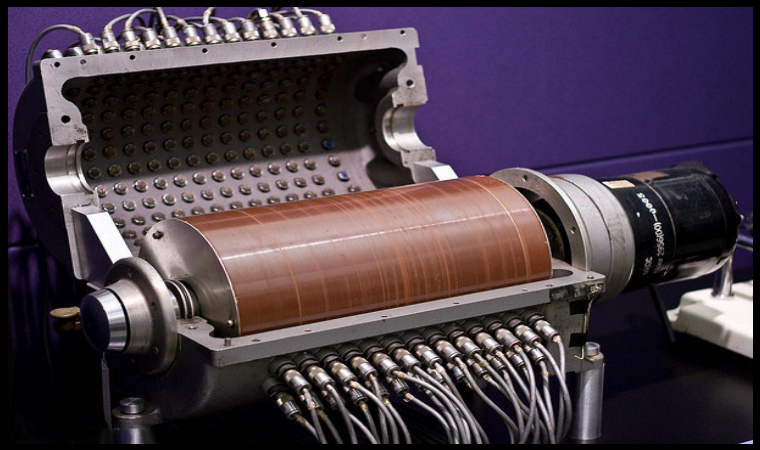


Core NVRAM (16kb)

IBM Mainframe room with Core Mem and Tapes as Input + cards/Output with Printer and to Tape

Evolved to Minicomputers & HDD

Digital Equipment Mini-Computer
With SRAM and DECtape for input

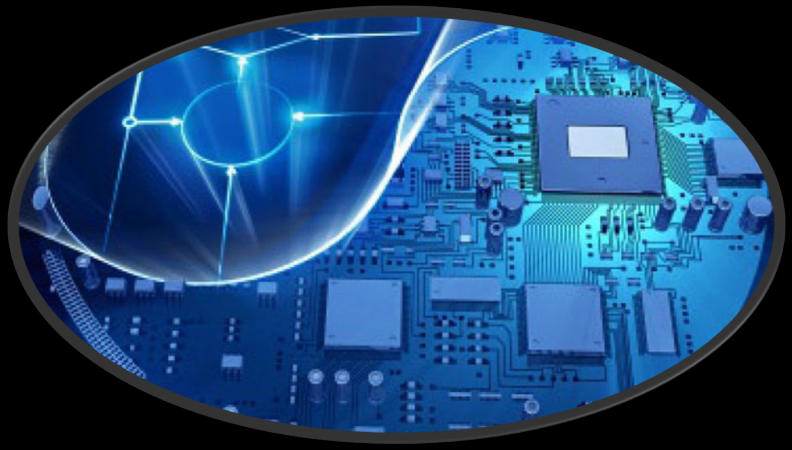


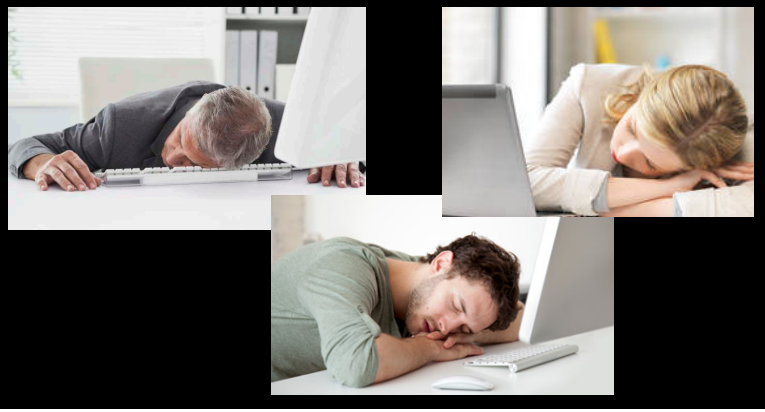
Magnetic
Drum
Drive

IBM HDD

Dreams of Software Engineers

- It has been the 'Holy Grail' for all software engineers and those making servers to have NVRAM instead of cache(DRAM), but have been living with a flawed Von Neumann architecture for almost 40 years!!
  - It has meant that external storage has been an intimate part of all applications over slower interfaces whether today's PCI or Ethernet adding to the time it takes to process data through an application!!
  - Sun made a statement before they were purchased by Oracle that external storage was dead as NVRAM was soon coming….not yet…
    - It has been about 12 years since then and we now see the re-birth of NVRAM with Intel's Optane, WD's MRAM, etc.., but today's NVDIMMs, which are a hybrid of RAM and flash are too early in my opinion to use as cost and density, let alone complexity of having firmware IN YOUR MEMORY is just one more thing to deal with from upgrades to testing.





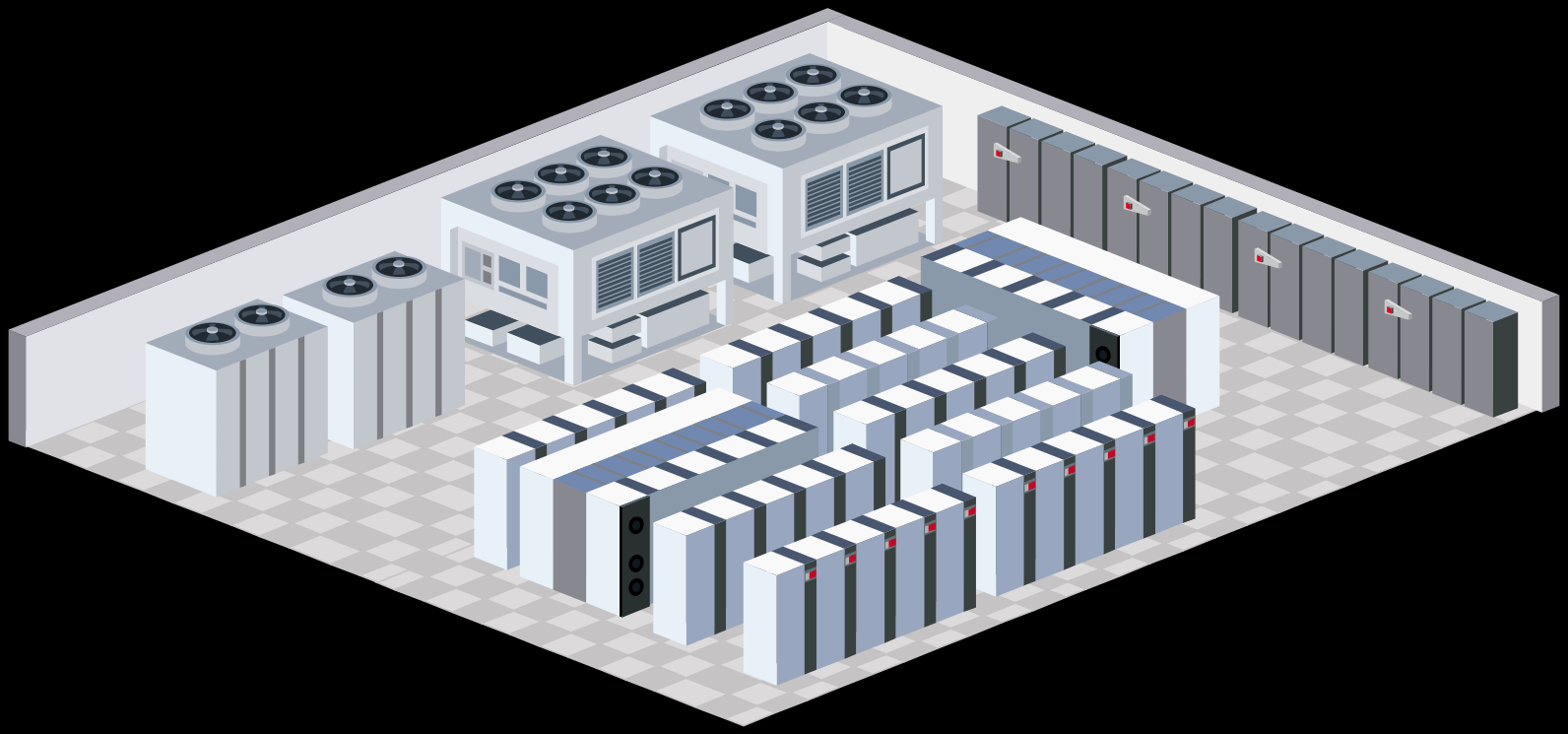
An NVRAM Server…what does it mean?

- The need for a NVRAM optimized Operating System that has all primary storage in memory, sharing with other NVRAM servers with same, making distributed computing much faster than waiting for peripheral storage latency to be endured.
- Memory Densities with Non-Volatile capability are reaching speed and have the density to allow for a ‘back to the future’ type of compute/storage relationship, circa 1960’s-1970’s ….
- As noted earlier, the software and data can co-exist inside the server, making it a full ‘computer’ as they were defined 50-60 yrs. ago.
- Peripheral storage is still vital, but Sequential data rate is more important than Random I/O
- Network Bandwidth is critical for access to storage for input data and output of application information, let alone distributed computing, given that memory bandwidth is over 1000x that of peripheral storage.

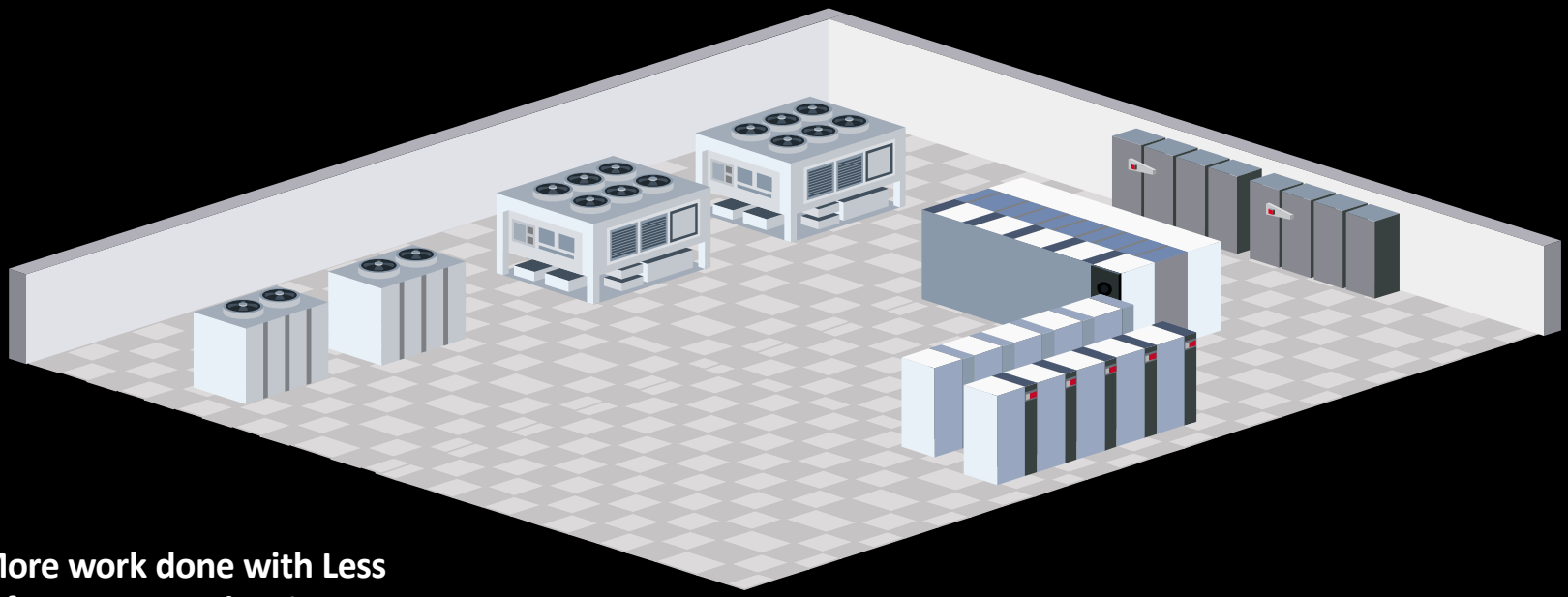
Implications of an NVRAM Server

- Applications and data volumes for application to work from, using peripheral storage only for input data as well as output of information (or backups).
- Distributed Computing ends up being magnitudes faster because of no random IO to peripheral storage
- NV Memory sizes today are capable of handling most working sets for many VM's per Server, fulfilling the dreams of software engineers for the last 4 decades!
- More work is done with less infrastructure…which means it leaves space for more work to be done with more of same, accelerating business like never before, especially for Big Data, AI, and all the other applications begging for from Databases, Mail, VDI, Financial software, you name it.

Datacenters Today



DataCenter with NVRAM Servers



- **More work done with Less infrastructure….leaving space to do magnitudes more work**

Summary

- NVRAM based servers, accompanied by a NVRAM optimized Operating System for transparent paradigm shift are just about here…this is a big tipping point…Performance numbers from 3 companies, show, depending on NVRAM type/size and CPU type a range of performance from 5-20MM IOPS @ 4-8K chunks!!!!
- Another tipping point is in network switching that reduce cost, like new/old networking (Rockport Networks Stealth CO.) will emerge that allows full cross sectional bandwidth and memory like hops between servers and storage, that an NVRAM server requires and at 1/2 the cost & power of today's switch-based networking.
- Ethernet based storage(NVMeOF) to assist in disaggregated server and storage is almost here (NVMeOF…allowing intelligent arrays(SDS or embedded) with cut-through performance BUT over ethernet as it should have been to start….back a while ago….

Summary

- Memory Fabrics as well (**e.g. GEN Z, etc.**) hold the promise to assist in a complementary way as faster networks for transparent Memory bandwidth across servers for better/faster distributed computing
- Servers that can handle more and more memory are coming….
- Servers with Gen4 PCI to allow the bandwidth between servers for distributed computing are coming…..ITS ALL COMING TOGETHER ENABLING ALL THE ABOVE!
- Evolution vs. revolution has always been tough, but SSD's were a revolution, so why not more?

Thank YOU!

Stephen J Sicola
Founder & CEO StorTrek
Consulting/www.stortrek.com
Stephen.Sicola@stortrek.com
Gumbydi@aol.com