SFD13 - X-IO Technologies (@XIOedge)

*X-IO Introduction - Gavin McLaughlin
Around for about 20 years
Intelligent Storage Unit (ISE)
1500 customers, more than 8000 units in the field

Q1 2016 Portfolio
100 Series - archive array
200 series - disk array
700 Series - hybrid array
800 Series - All Flash array

iglu blaze
full-featured SAN array
distributed controller architecture
modular scalability

Don't want to turn back on existing customers
The market is changing
Revenues increasing, net losses increasing as well - "The madness had to stop"

Strategy Pivot
EBITDA positive
Eliminated all debt
Raised new capital

From -> To
From
Midrange enterprise storage systems
- innovate storage firmware IP
- Exceptional engineering then
- custom hardware with difficult supply chain
- Struggling in hyper-competitive market
- new customer acquisition cost unsustainable

To
Edge computing systems
- real-time big data analytics computing
- converged server and storage
- better, simpler, cost effective PCIe plus NVMe alternative to Dell EMC DSSD
- Dense packaging for edge applications
- Infrastructure software to provide edge data protection and management
- Leverage software defined NVMe flash storage IP
Next generation SSD storage systems
- next generation storage systems for existing X-IO storage customers
- leverage fast reliable ISE firmware with commodity off the shelf ODM hardware
- innovation in data deduplication and add native data services

ISE G4

- next generation storage for existing X-IO storage customers
- leverages fast reliable ISE firmware with standardised, open datapac and systems hardware for lower cost and highest performance
- native data services including patented faster deduplication technology

Axellio
Converged server and storage for the big data real time edge
Unique FabricXpress large PCIe + NVMe fabric architecture
High performance analytics processing with 2x2 CPU Intel server motherboards
Large, persistent, in-server NVMe data store (up to 1PB)
High bandwidth data ingestion (>200Gbs or >30GBs)
Simultaneous ingestion and random access processing of stored data (480Gbs or 60GBs full duplex at <50micro-seconds average access latency)

*Edge Computing - Bill Miller
Too late big data analytics
Batch Approach
Sophisticated data science methods
big working data sets
results come later
based on data that is old
Insight/response in hours, days, weeks, months

Too Small Fast Data Analytics
Streaming approach
Small data sets that fit in memory
Simplistic analysis methods, limited insights
Results come fast
Based on data that includes new data
Insight/response in milliseconds, seconds, minutes

Big Fast Data
Big working data sets
Include new data as it is created
Sophisticated big data analysis and AI methods
Real time insights that enable real time response

The real-time big data edge is where:
The time that would be needed to move the data to another place for analysis would result in an unacceptable response time
Therefore, the data must be ingested and analysed at its place of origin to generate a real time response
These problems are exacerbated when the rate of data creation is very high, large data sets are required for analysis, and the time to isight and action requirement is very short

Cyber security
Intelligence
Trading surveillance

Real Time Big Streaming Data Analytics (photo)

*Addressing Edge Computing with Axellio (David “Gus” Gustavsson)
Why build a different platform in today’s market?
VM oriented cloud services drive a “HW agnostic” mindset in the industry • “I don’t care what HW it runs on as long as the SW does the job”
• Cloudscale architectures are optimizing for providing services for many customers at large scale. • Cheapest possible HW at large scale automation
• Server vendors have focused many of their offerings towards scale out architectures rather than extending the in-box capabilities
• Big-data analytics is driving a need to ingest and analyze very large amounts of data fast
• Hadoop is parallelizing the problem and spreads it over MANY nodes
• One system that can do the data ingest/traversal of 30 “normal servers” can dramatically change things

Design objectives and principles
Design objective
System platform with next generation storage capabilities
- breakthrough IO performance
- breakthrough storage capacity density
Guiding principles
Drive cost efficiency by using off-the-shelf components and technologies
Package into a small form factor for density efficiencies

What is the edge?
Data has mass
Once the dataset becomes big or fast enough, it’s easier to move the application to the data than vice versa
This is the Edge
This also means you need appropriate CPU/RAM to run the application attached to the large dataset
• Uncork the I/O bottleneck and you can now use a lot more CPU/RAM to do something useful

Where does Axellio fit?
Embarrassingly parallel or inherently serial
Different problems require or are better served by different architectures
• Some problems are completely independent and can easily be parallelized
• Other problems have a varying degree of serialization due to data dependency • The higher the data dependency the harder it is to scale out

Scale-out - “Built for large scale” (photo)

Scale up - “We need a bigger boat”
• Grow the server capability to hold the entire or more of the application
1. The problem is inherently serial
• Cannot be split to run on different nodes

• Iterative numerical methods, 3-body problem
2. Data ingest requires access to more and faster data than is held in one node
• Problem can be split but access to the data becomes the inhibitor
• Traditionally used SANs with shared external storage
• With hyper-converged, architectures the network carries the load
3. Data dependencies between tasks becomes the bottleneck
• Node to node network latencies/BW inhibit performance
• Answer today: Build a faster network!
• Maybe there is a faster and cheaper way.

Cost effective performance, Ethernet vs PCIe
Ethernet vs PCIe? - Apples and helicopters? • External network vs system bus?
As a communication media PCIe is cheaper, faster and more efficient than Ethernet
• Each Xeon E5-26xx v4 CPU is equipped with 40 PCIe Gen 3 Lanes => ~40GB/s @ Half Duplex
• Even the E5-2603v4 @ $213* has 40 lanes built in waiting to be utilized at no extra cost
• Bare wires, no extra protocol conversion HW
• High speed signal integrity limits practical distance
• Use it as intended, very fast internal bus to move a lot of data fast and cheap.
Scale out vs scale up (scale in).
• Scale out is not bad, maybe not the most cost effective
• Scale in before scaling out - Get on the Bus!
• *https://ark.intel.com/products/92993/Intel-Xeon-Processor-E5-2603-v4-15M-Cache-1_70-GHz

Key design decisions
NVMe over PCIe instead of 12Gb SAS
- a more efficient protocol
- early enough that dual ported NVMe SSDs did not yet exist
- Worked with SSD manufacturer to help test and develop the dual port capability
In house motherboard => Intel Xeon motherboards
- standard form factor, cost effective, volume production
- great flexibility of compute capacity to match the I/O capability, up to 44 cores & 1TB RAM per motherboard
72 dual ported NVMe SSDs in 2RU
- provides a large capacity pool as well as many devices to spread the I/O load over
Modularity in design for offload modules / SSD carriers
- provide flexibility for different use cases
Networking capability to also enable high speed scale out architectures where needed

Axellio - Technical Specifications
[X-IO_Axellio_Spec.png]
2RU form factor
4 socket Intel e5-26xx v4 CPUs
- 16 to 88 cores and 24 to 176 threads
- core optimised or frequency optimised
32 DIMMs, 16GB - 2TB
- optional NVDIMMs for storage cache

Enterprise Grade, Industry-Standard NVMe Storage
- Up to 72x 2.5” NVMe SSDs
- 460TB of NVMe Flash with 6.4TB NVMe SSDs (1PB coming)
- >12 Million IOPS, as low as 35 microseconds latency, 60GBs sustained
- Octane ready
Optional offload modules
- 2x Intel Phi - CPU extension for parallel compute
- 2x Nvidia K2 GPU - Video processing, VDI
- 2x Nvidia P100 Tesla - Sci Comp, Machine Learning
- Solarflare Precision Timing Protocol (PTP) Packet Capture (PCAP) offload

Axellio System Architecture and Capabilities
PCIe / NVMe Primer
PCIe: (https://pcisig.com/)
• Peripheral Component Interconnect Express • Introduced Gen 1 2004
• Generic serial point to point expansion bus
• Full duplex links between endpoints
• Links can have 1-16 lanes (x1,x4,x8,x16)
• 8 GT => 125 picosecond bit timing
• Each Gen 3 Lane => 985 MB/s in each direction.
• x16 link => 31.5 GB/s full duplex
• Xeon E5-26xxV4 CPUs all come with 40 PCIe lanes

NVMe: (http://www.nvmexpress.org)
• Device access protocol designed bottoms up for low latency media over PCIe
• First spec 2011
• Many (64k) deep (64k) queues enables lockfree
I/O across many CPU cores in parallel
• <1/2 CPU cycles per I/O than SAS/SATA
• Lowers “Software latency”
• Frees up CPU cycles for other work
• Leaner protocol for drastically faster I/O

FabricXpress
Extends the native PCIe bus significantly
PCIe based Interconnect
- Up to 72 NVMe SSDs - significantly more SSDs
- between server modules
- offload modules
Dual ported NVMe architecture
- allows access to same data on same SSD from both servers
- Shared access for HA solutions
- Enables independent server behaviour on shared data

Networking and offloading module
Networking
- 1x16 PCIe per server module for networking
- supports standard off the shelf NICs/HCAs/HBAs
- supports HHHL or FHHL cards
- Ethernet, InfiniBand, FC

- Up to 2x100GbE per module
Offloading Module
- Two centre modules is replaced with single carrier
- Holds two FHFL DW, x16 PCIe cards
- Nvidia P100: +18.6 Teraflops (sp)
- Nvidia V100: +30 Teraflops (sp)

Edge data analytics platform
Ingest and analyze data at unprecedented speeds
2U replaces a rack of scale out gear
Uniquely qualified for real-time big data analytics
[X-IO_Edge_data_analytics.png]

Edge Computing recap
• Converged server and storage for the big data real time edge
• Unique FabricXpress large PCIe + NVMe fabric architecture
• High performance analytics processing with 2x2 CPU Intel server motherboards
• Large, persistent, in-server NVMe data store (up to 1PB)
• High bandwidth data ingestion (>200Gbps or >30GBps)
• Simultaneous ingestion and random access processing of stored data (480Gbps or 60GBps full duplex at <50µs average access latency)
• Capable of next-generation storage platform through ISE software

*Improving deduplication with mathematics (Richard Lary, CTO)
Deduplication requires significant work (slide)

X-IO Approach to Deduplication