TEKHEAD
IT'S ALL ABOUT THE WAF

NODDY'S ~~GUIDE~~ Warty TO STORAGE DESIGN

Storage 101

Alex Galbraith
vExpert 2013/14, VCAP4/5-DCD, VCP3/4/5
@alexgalbraith
www.tekhead.org

QUICK SURVEY

- Do you manage or design storage on at least a regular basis?
- Is storage your primary role?

TEKHEAD
IT'S ALL ABOUT THE WAF

NODDY'S GUIDE TO STORAGE DESIGN

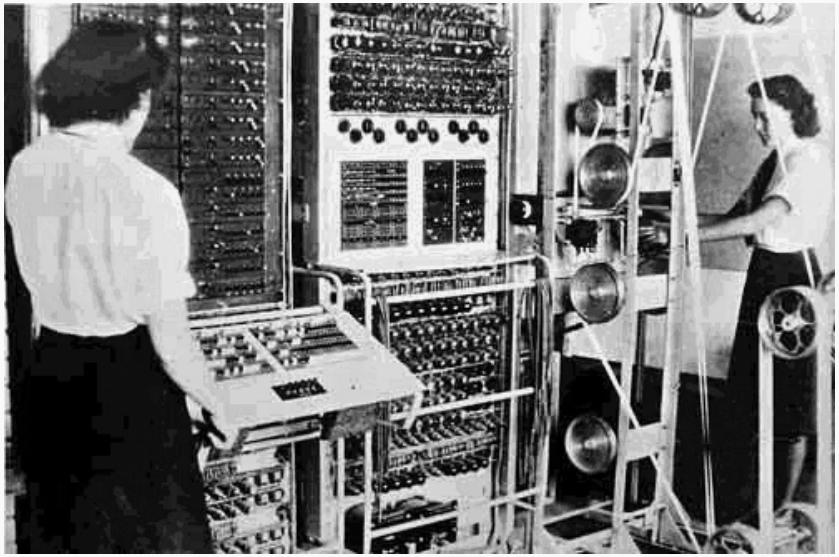
"Storage is Boring!"

Anil Valluri

President, India & SAARC at NetApp

A BRIEF HISTORY OF STORAGE…

- Ye Olden Days…
  - 1920s – Magnetic tape patented
  - 1930s – Magnetic drums
  - 1940s – Cathode-ray "storage" tubes
  - 1950s – Hard disks
    - 50x 24" platters = 5MB



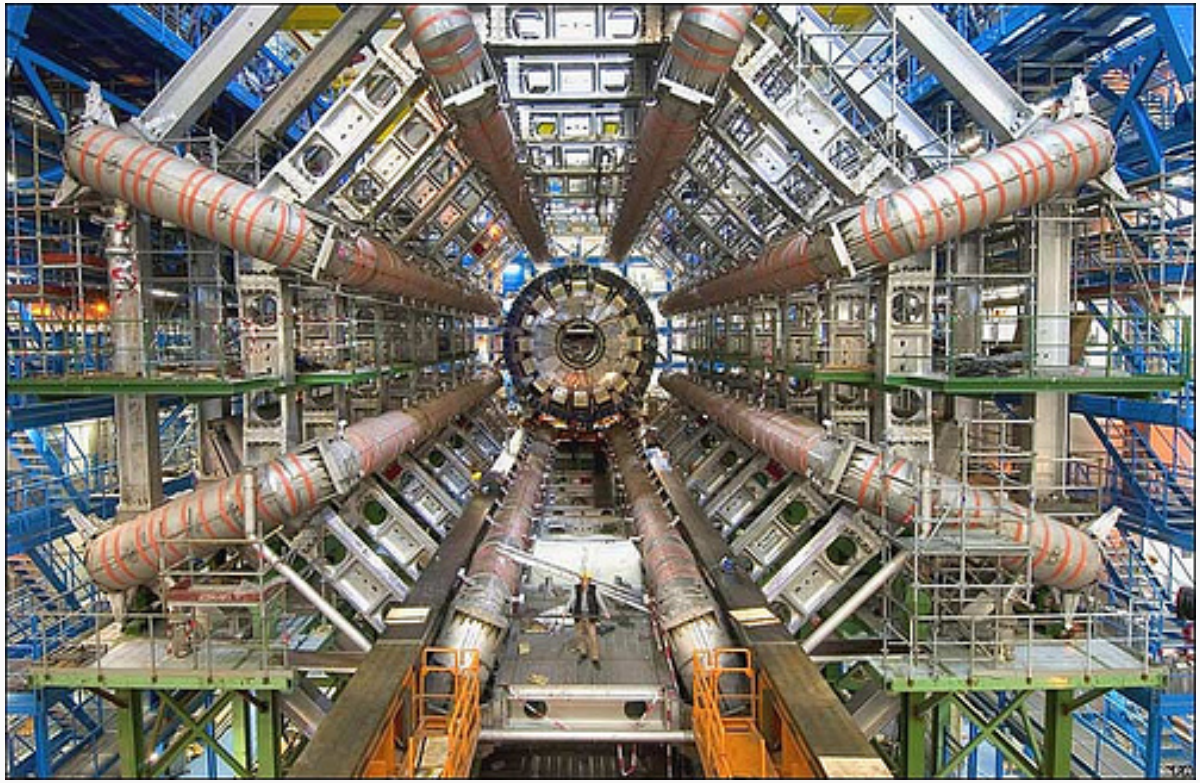
TEKHEAD
IT'S ALL ABOUT THE WAF

A BRIEF HISTORY OF STORAGE…

- 1980s-2000s
  - SCSI 1/2/3
  - Ultra SCSI 2/3(160)/320
  - SATA / SAS

TEKHEAD
IT'S ALL ABOUT THE WAF

A BRIEF HISTORY OF STORAGE…



- Today 4TB drives – same number of IOPS
  - (And tape!)

TEKHEAD
IT'S ALL ABOUT THE WAF

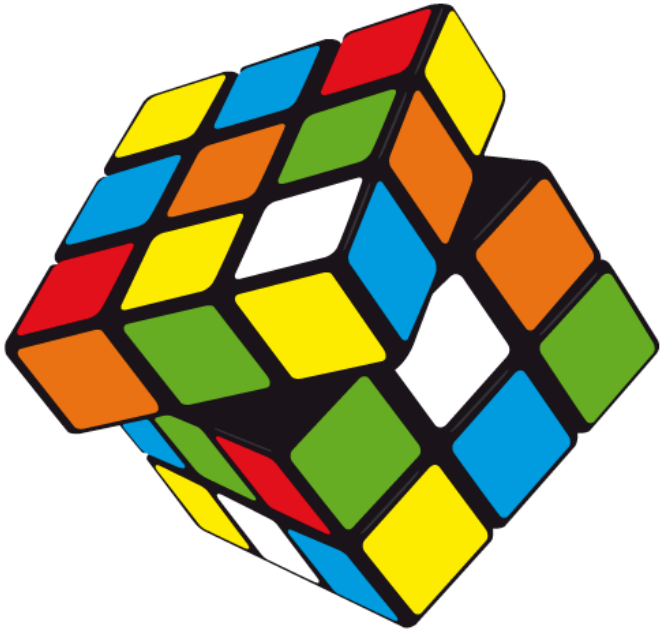
FLASH!

- All purpose Flash!
- Invented ~1980 but stupidly expensive
- Not just high IOPS, also very low latency
- Options at most layers in the storage path
  - Host
  - Cache
  - Primary Storage (T0)
  - Standard Storage (T1)
- Often used as a “band aid” for badly written apps
- Like hard to reach stains, flash doesn’t solve every problem…
  - Still limited by array backplane speeds
  - Sequential perf can be no better than spindles
  - Optimum perf requires free space
  - Drives have limited maximum writes (getting better)
  - Quite expensive, but getting cheaper by the day!

TEKHEAD
IT'S ALL ABOUT THE WAF

DESIGN CHALLENGES

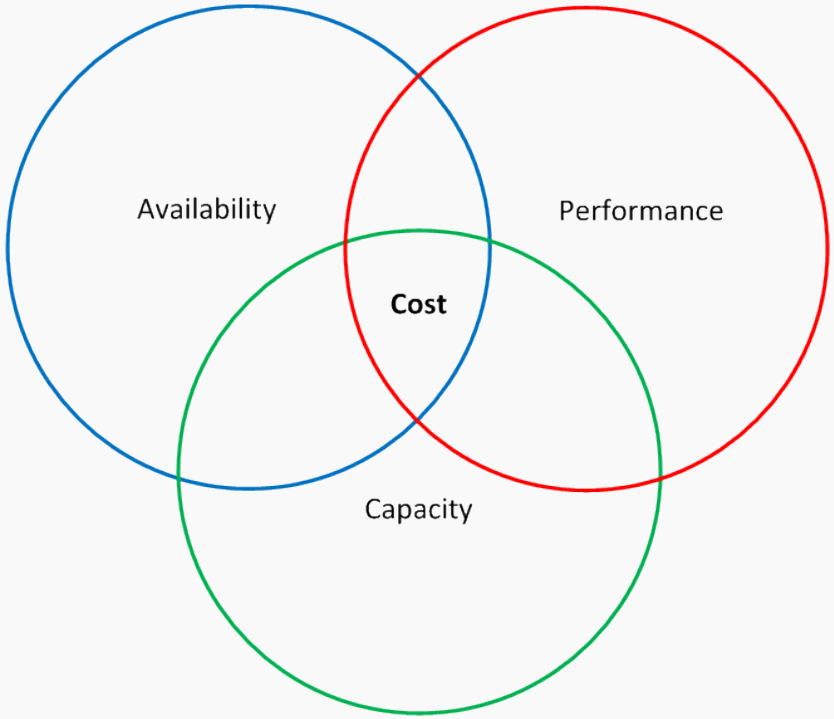
- So many choices!
  - "Where do I place me bets?"
- Complex terminology
- Storage performance
  - IOPS!
  - Throughput (to a lesser extent)
  - Better informed customers & slick marketing!
  - SLAs / OLAs used to be about availability & capacity
- VM and storage sprawl
- VM Sizing?
- High Availability
- Shrinking budgets
  - "We'll just have to do more with less"



TEKHEAD
IT'S ALL ABOUT THE WAF

DESIGN CHALLENGES

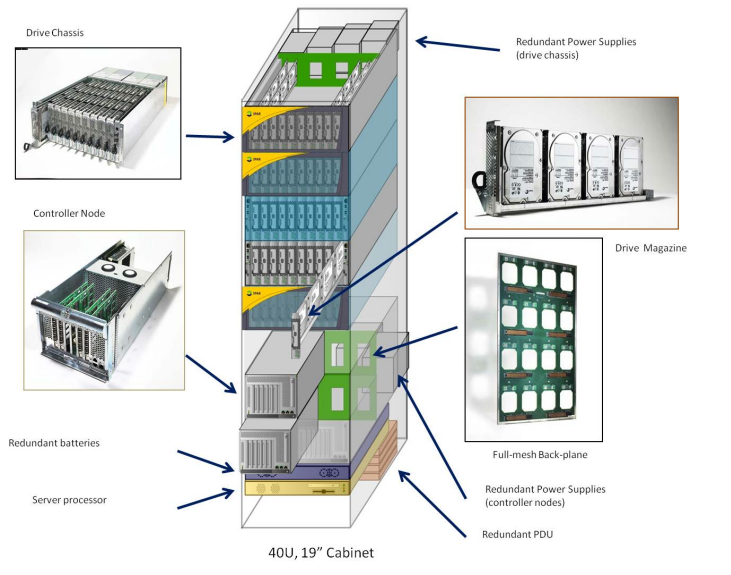
- Design is a balance:



TEKHEAD
IT'S ALL ABOUT THE WAF

ARRAY RESILIENCE

- Redundancy and SPOFs
  - Always use dual head arrays as a minimum
  - Dual heads does not guarantee resilience!
    - Replicate critical data across arrays
  - Check your vendor docs for SP down scenarios
    - Some go into write-through mode
  - Shelves can and will fail!
    - Always RAID across shelves if possible for HA

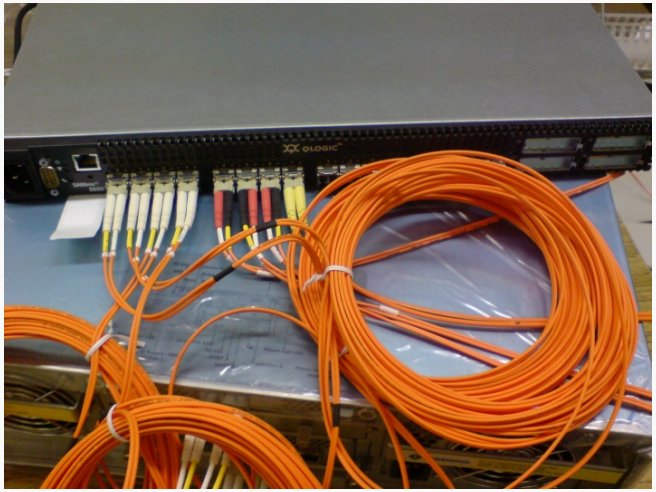


- IO Path
  - Front End Ports / HBAs
    - Service Processors / Controllers / Heads / Nodes
      - Switches / backplane
        - Trays / Shelves / Drive Chassis
          - Disks (Spindles / flash)

TEKHEAD
IT'S ALL ABOUT THE WAF

WHAT'S YOUR FAVOURITE FLAVOUR?

- Block
  - Fibre Channel
    - Very low latency & lossless
    - More secure as dedicated and non-routable
    - More Complex
    - Can be Expensive (Switches / HBAs / Licensing)
    - Supports RDMs & MSCS
  - iSCSI
    - Easy to implement
    - Relatively inexpensive
    - Fewer cables & switches so less power (Use NetIOC)
    - Supports RDMs
    - Less secure than fibre channel
    - MSCS – In guest
    - Slow path failovers compared to FC (increase guest SCSI timeouts)
  - FCoE
    - Just a stop gap!

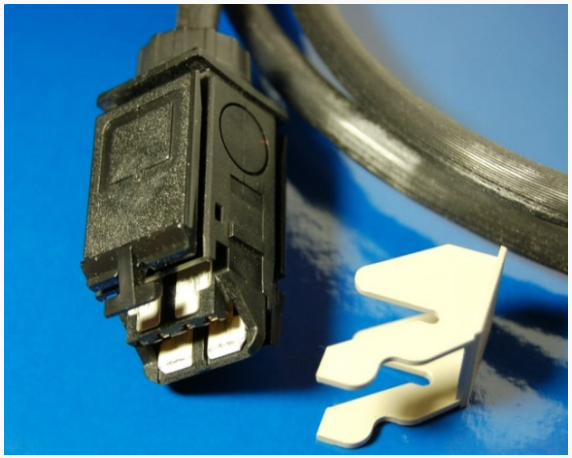


TEKHEAD
IT'S ALL ABOUT THE WAF

WHAT'S YOUR FAVOURITE FLAVOUR?

- File
  - NFS
    - Very easy to implement (esp for Linux admins)
    - Relatively inexpensive
    - Less secure than fibre channel
    - Fewer cables & switches so less power (Use NetIOC)
    - No MSCS
  - SMB
    - Hyper-V
    - May see more adoption with SoFS and CSVs

- FCoTR
  - Fibre Channel over Token Ring
    - http://blog.fosketts.net/2010/07/16/fibre-channel-token-ring-fcotr/



TEKHEAD
IT'S ALL ABOUT THE WAF

WOULD YOUR STORAGE VENDOR EVER LIE TO YOU?! (PT. 1 - CAPACITY)

- Know your GB and TB from your GiB and TiB!
  - Base 10 (Decimal) vs Base 2 (Binary)
  - Formatted drives are in Base 2

Decimal	Metric	Bytes	Binary	Metric	Bytes	Multiplier
byte	B	1	byte	B	1	1
kilobyte	kB	1,000	kibibyte	KiB	1,024	0.977
megabyte	MB	1,000,000	mibibyte	MiB	1,048,576	0.954
gigabyte	GB	1,000,000,000	gibibyte	GiB	1,073,741,824	0.931
terabyte	TB	1,000,000,000,000	tebibyte	TiB	1,099,511,627,776	0.909
petabyte	PB	1,000,000,000,000,000	pebibyte	PiB	1,125,899,906,842,624	0.888

IOPS

- Read vs write ratios (e.g. 50/50, 80/20)
- Faster spindle speeds = more IOPS
- Always err on the safe side
- Typical IOPS per spindle type:
  - SSD 6000+
  - 15K FC/SAS 175
  - 10K FC/SAS 125
  - 7200 SATA/NL SAS 75
- Random vs sequential IO?
  - Transactional vs Analytical
  - VM traffic tends towards random (IO blender)
- Real world examples:
  - SharePoint – ~0.5 IOPS/*GiB*
  - Exchange (Cached Mode) – 0.12 IOPS/*Mailbox* + replication
  - Exchange (Online Mode) – Depends on the mailbox size!
  - Information Worker File Storage – 1.5 IOPS/*User*
  - Redirected profiles – 1.5 IOPS/*User*
  - Session desktops – 3/6/12 *IOPS/User*

TEKHEAD
IT'S ALL ABOUT THE WAF

WOULD YOUR STORAGE VENDOR EVER LIE TO YOU?! (PT2 - PERFORMANCE)

- My IOPS are bigger than yours!
  - Most (not all) vendors use 4k IOPS @ 100% read
  - Few real-world IO profiles are 4k IOPS
  - Some vendors state max IOPS under very specific conditions (e.g. 100% local access per node)
  - Different vendors report differently on IOPS
    - Key for providing SLA reporting
    - Normalisation of statistics can be very useful
  - Many never mention latency…
- Always ask your vendor for an estimate based on your expected workload

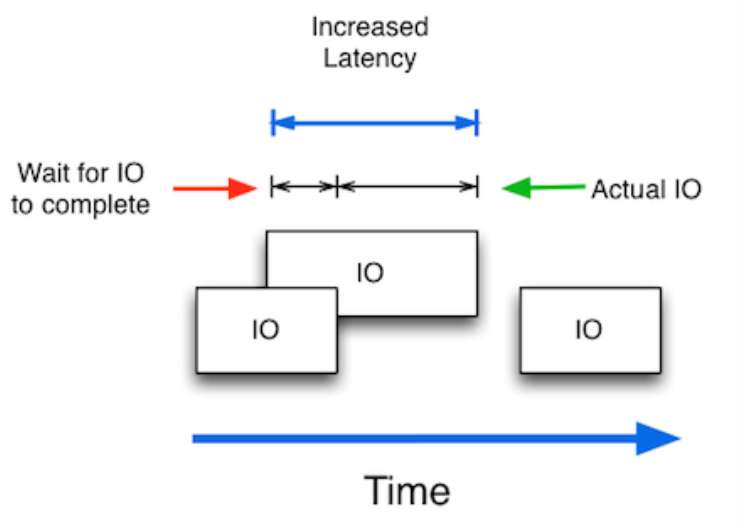
TEKHEAD
IT'S ALL ABOUT THE WAF

L… A.. T… E… N. C……Y!

- The Mr Hyde to IOPS Dr Jekyll
- Sometimes referred to as:
  - Response Time
  - Round Trip Time
- Total time for one IO

- *More latency = more user complaints!*

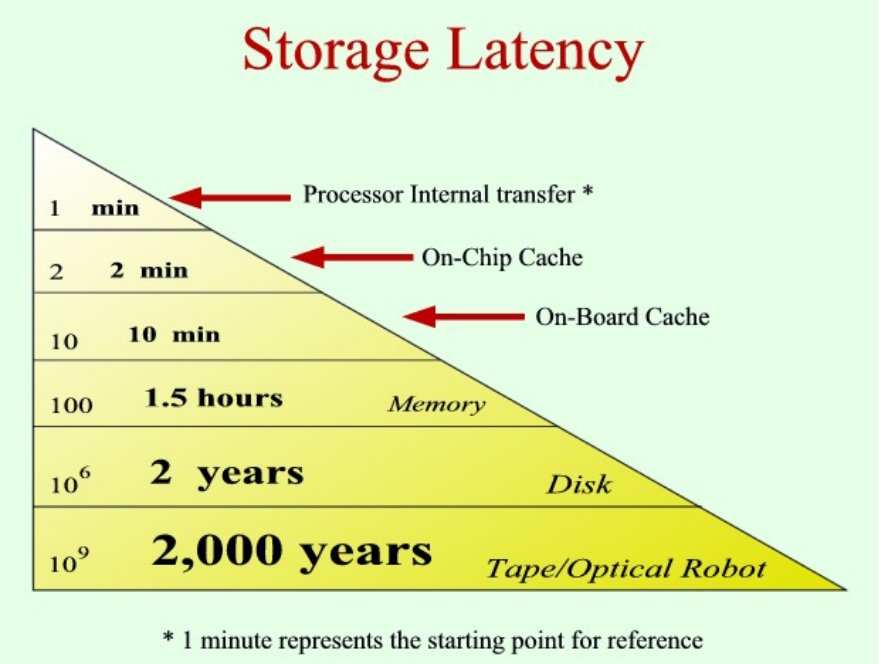
- Example vendor array
  - 39k IOPS @ 30ms
  - 25k IOPS @ 12ms



TEKHEAD
IT'S ALL ABOUT THE WAF

L… A.. T… E… N. C……Y!

- Comparative Latency:

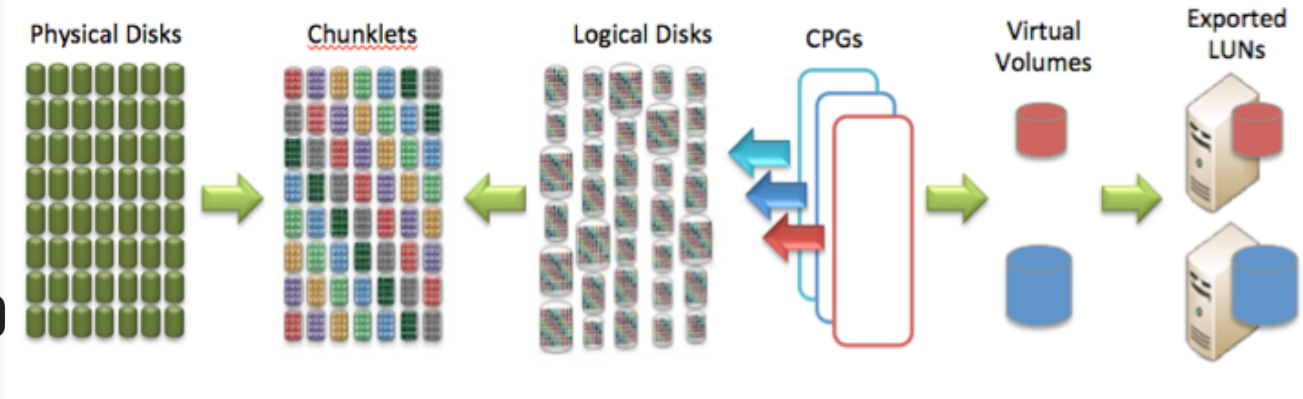


- Typical acceptable workloads / tiers:
  - High performance <2ms
  - Standard 10-15ms
  - Archive / low tier 30ms+

TEKHEAD
IT'S ALL ABOUT THE WAF

WIDE STRIPING

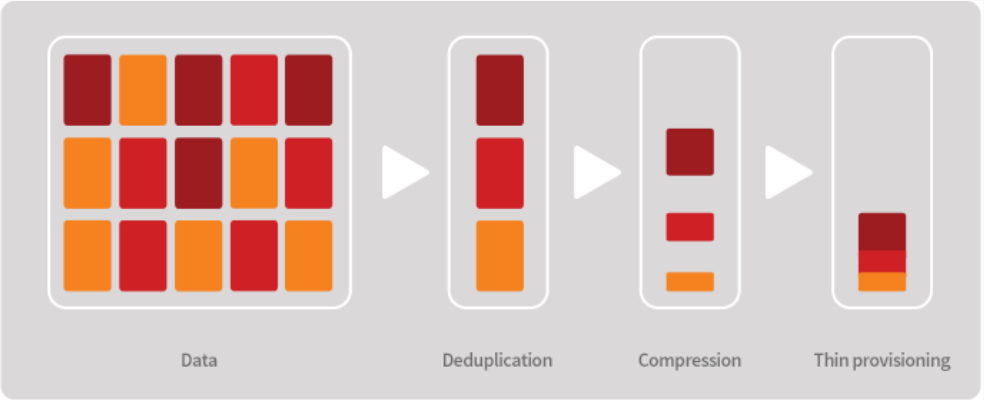
- **WARNING! Opinion Inbound!**
- Are those 256mb chunklets in your pocket or are you just pleased to see me?
- Pros
  - Don't waste IO
  - Better average performance for all VMs
- Cons
  - "Noisy Neighbours" can be a nightmare
- Recommendation: Only use wide striping if you are using QoS, e.g.
  - 3PAR Priority Optimisation
  - SIOC (if 100% virtual)



TEKHEAD
IT'S ALL ABOUT THE WAF

DEDUPLICATION & COMPRESSION

- Dedup
  - File or block level
  - AFA vendors need dedup to be competitive £/GB (E.g. Pure / SolidFire)
  - Some only dedup SSD tiers not spindle
  - Some data types dedup better than others
    - Database – Poor
    - Media – Poor
    - File – Good
    - VDI – Excellent
  - Encryption kills dedup
    - You cant dedup random data!
  - Block / Page size can have significant impact on dedup levels & performance (e.g. 4/8/16k)



TEKHEAD
IT'S ALL ABOUT THE WAF

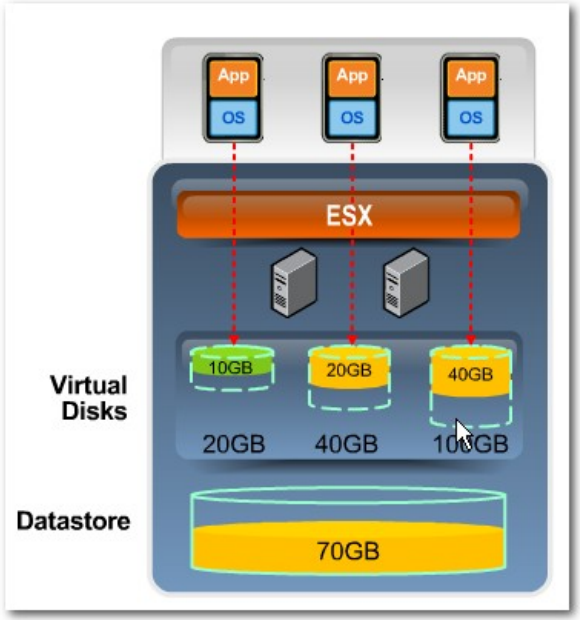
DEDUPLICATION & COMPRESSION

- Compression
  - Typically more effective than dedup
  - Very poor on media (video / photos)
- How do I estimate my compression?
- Inline vs post-processing
  - Post can risk performance and capacity
  - Can be dependent on load!
- ASICs vs Software
  - SDDC = Software Defined Dedup and Compression?

TEKHEAD
IT'S ALL ABOUT THE WAF

THIN PROVISIONING

- Can save a fortune in storage costs!
  - More than 2x in some environments
- Can be done at:
  - vSphere layer
    - Thin
    - Thick vs Eager-Zero Thick
  - Storage Layer
    - Thin provisioned block storage
    - NFS is built in
  - Both – Not recommended



Remember!

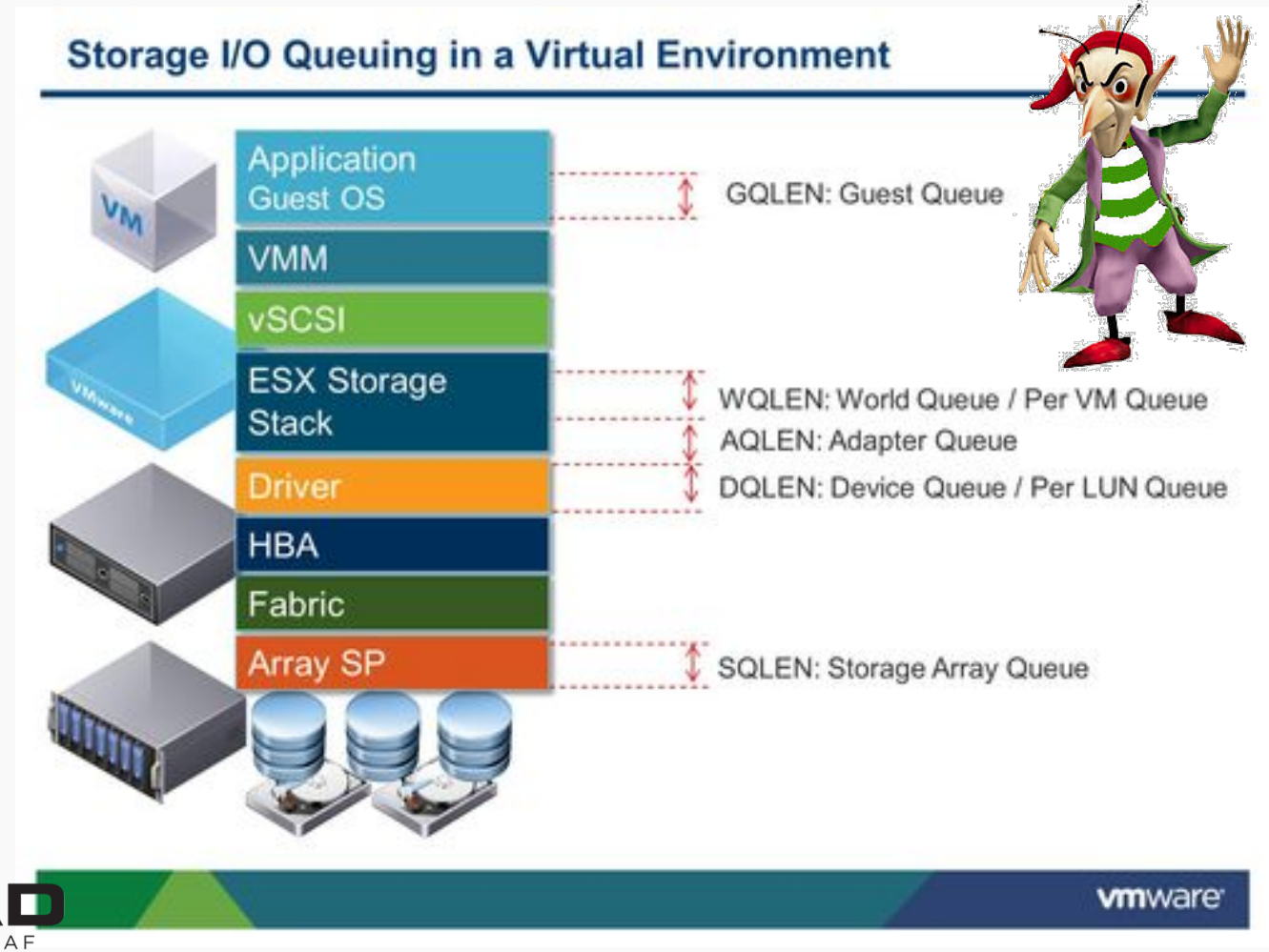
Dedup, Compression & Thin Provisioning techniques *reduce effective IOPS/GiB*

TEKHEAD
IT'S ALL ABOUT THE WAF

QUEUE DEPTHS

HERE BE GREMLINS...

- I strongly suggest you read up on this!
- Fan-Out Ratios
- FC vs iSCSI



TEKHEAD
IT'S ALL ABOUT THE WAF

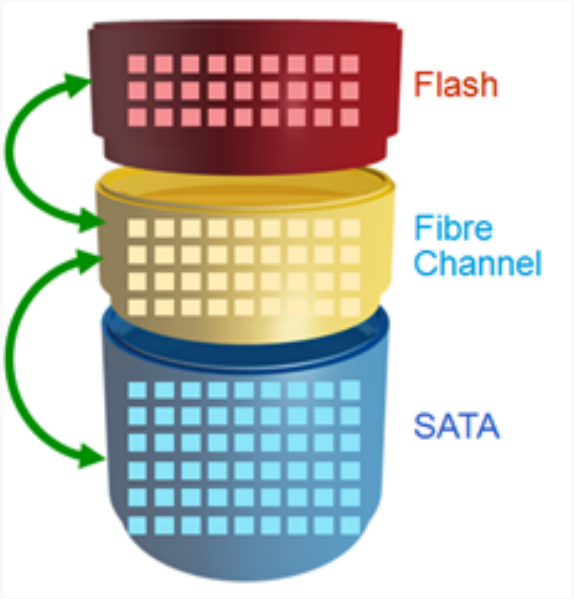
QUALITY OF SERVICE

- Most mid range arrays and above have some form or another
- Helps mitigate negatives of Wide Striping
- Less effective on spindle systems
- May attract additional license fees!
- Typically 2 Types:
  - Soft
    - Best efforts to meet minimum and maximum
  - Hard
    - Guaranteed minimum and maximum IOPS
    - More typical on AFAs

TEKHEAD
IT'S ALL ABOUT THE WAF

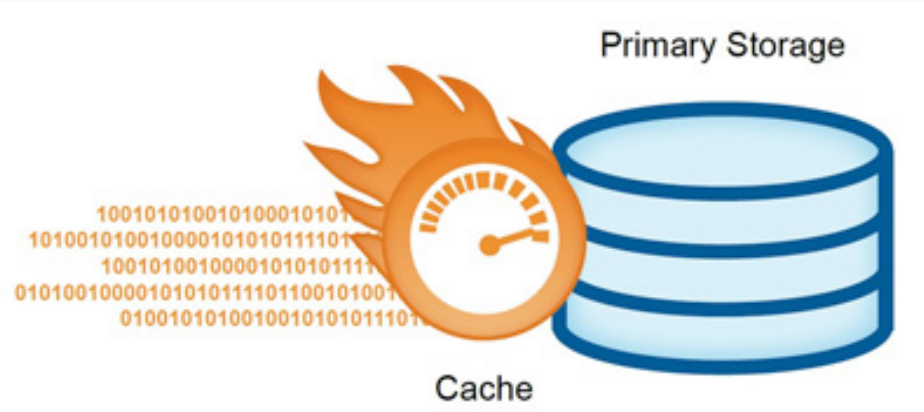
AUTO-TIERING

- The clue is in the name!
- Can be different tiers of same storage tech
- The storage challenge:
  - 70% of data is static and untouched
  - 5% of data generates 60% of all IO
  - Remaining 25% generates some IO
- Can save a fortune by using auto-tiering techniques with no perceived performance loss
- Only appropriate for some workloads
- Peaky workloads can have poor performance
- Can pin some workloads, but then its manual auto-tiering?
- Waterfall & real-time tiering can help



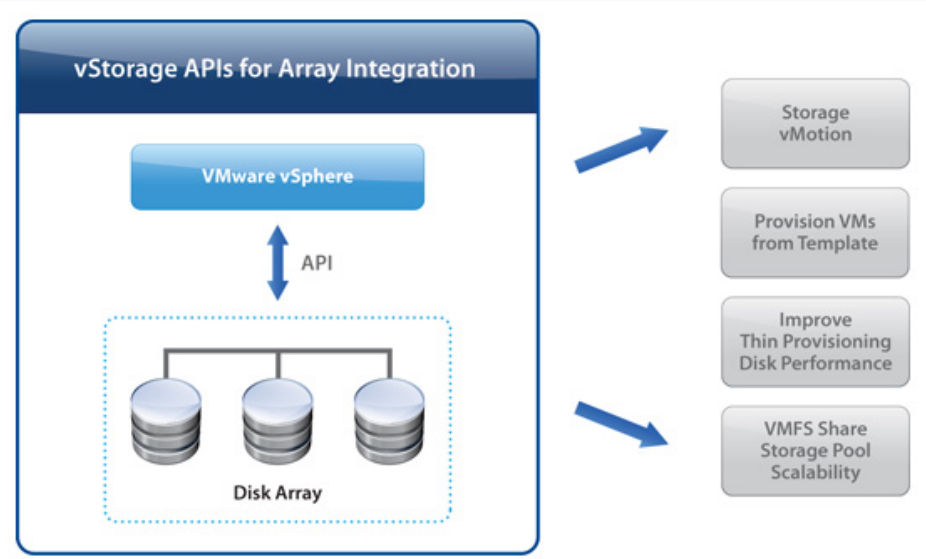
CACHING

- More is always better!
  - Varies from 256MiB to 100+ GiB
- Cache options at every level:
  - Hypervisor
    - e.g. Fusion IO / Pernix Data / vFRC
  - HBA (local / SAS attached)
  - Array NVRAM
  - Array SSD
- Cache deduplication
- Read vs write
- Write-through vs write back?
  - Write Coalescing
    - NVRAM & SSD
    - De-staging to disk



VAAI

- VAAI
  - VMware vSphere Storage APIs for Array Integration
  - Example Primitives (5.x+):
    - Full Copy
    - Block Zeroing
    - Hardware Assisted Locking (ATS)
    - Block delete (SCSI UNMAP)
    - Native snapshot offload
      - Doesn't speed up RAM copy for snaps
  - ATS does not mean you want giant datastores!



TEKHEAD
IT'S ALL ABOUT THE WAF

DESIGN RULES OF THUMB

- Performance
  - Design for performance first, capacity second!
  - Ask your vendor for a copy of their sizing calculator and play with the numbers
  - Ignore cache in sizing calculations
  - Design for peak performance, not average.
  - Flash, aaaah!
  - Benchmarking the existing estates
    - Capacity Planner
    - 3rd Party Tools (Dpack, Platespin)
- Capacity
  - Use vSphere 5 to exceed 2TiB per VMDK
  - Keep minimum 10-20% free for overheads
    - VM Swap Space
    - Snapshots
    - Templates & ISOs
  - Thin provisioned arrays mean you can afford to overprovision datastores
  - Does your vendor support:
    - Deduplication
    - Compression
    - Auto-tiering
    - Thin Provisioning

TEKHEAD
IT'S ALL ABOUT THE WAF

DESIGN RULES OF THUMB

- Availability
  - High Availability at the OS & Application Layer
    - VM Clustering across hosts requires RDMs or guest iSCSI
- Cost
  - What is your key metric?
    - IOPS/GiB, £/GiB, £/IOPS
  - For 100% support always use vendor SFPs. To save a fortune consider 3rd party ones…
  - Cost inc Licensing, cabling, SFPs, etc
  - Rack Space / Power / Cooling
  - If budget allows, buy some growth up front for max discount
    - Except potentially AFAs
- Don't always assume you have to buy a new array!
  - Atlantis ILIO
  - PernixData
  - VSAN
  - Nutanix / Simplivity etc
  - SOFS

TEKHEAD
IT'S ALL ABOUT THE WAF

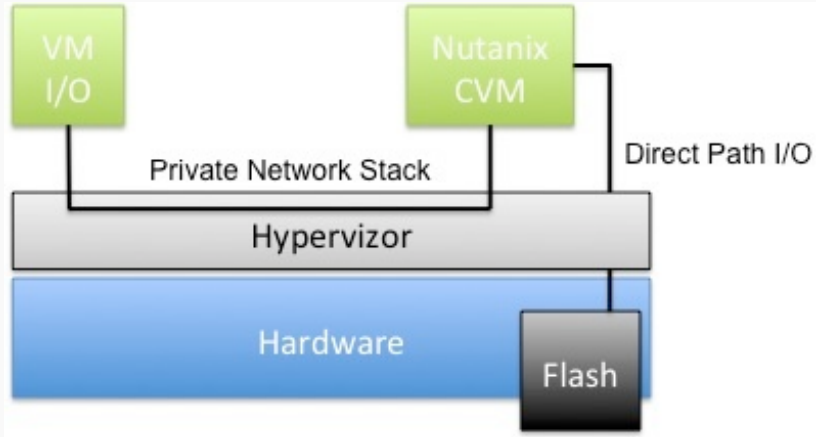
DESIGN DECISIONS AND IMPACTS

- Consider Impact of every design decision

TEKHEAD
IT'S ALL ABOUT THE WAF

FEELING HYPER-CONVERGED?

- Converged Systems (turnkey) e.g.:
  - VCE vBlock (Cisco UCS)
  - Flexpod (Cisco + NetApp)
  - HP ConvergedSystem
- Hyper-converged systems & VSAs:
  - HP StorVirtual VSA (Lefthand)
  - VSAN
  - Nutanix
  - Simplivity
- Capacity and compute requirements must scale roughly in line for Nutanix / Simplivity to be viable
  - VSAN is much more flexible
  - Dell XC could be interesting!



TEKHEAD
IT'S ALL ABOUT THE WAF

LINKS

- **Links –**
  - Great storage IOPS calculator
    - http://www.wmarow.com/strcalc/
  - UCS and HP Virtual Connect
    - http://www.wooditwork.com/wp-content/uploads/2012/11/Julian-Wood-vSphere-Networking-and-Converged-IO-with-Blade-Servers.pdf
  - Community Adapter Queue Depth List
    - http://www.virtuallyghetto.com/2014/06/community-vsan-storage-controller-queue-depth-list.html
    - https://docs.google.com/spreadsheets/d/1FHnGAHdQdCbmNJMyze-bmpTZ3cMjKrwLtda1Ry32bAQ/edit?pli=1
  - Troubleshooting Storage Performance in vSphere – Storage Queues
    - http://blogs.vmware.com/vsphere/2012/07/troubleshooting-storage-performance-in-vsphere-part-5-storage-queues.html
  - Mid-range storage array guide
    - http://www.dcig.com/2013/10/dcig-2014-enterprise-midrange-array-buyers-guide-now-available.html
  - #vBrownbag – VCAP5-DCD Storage Design
    - http://professionalvmware.com/2012/02/apac-brownbag-vcap-dcd-storage-follow-up/
  - #vBrownbag - Basic Storage Maths by Alastair Cooke
    - http://www.digitalpodcast.com/items/7680389 / https://itunes.apple.com/gb/podcast/professionalvmware-vbrownbag/id468638808?mt=2#
- **Books –**
  - VMware vSphere Design; Forbes Guthrie, Scott Lowe & Kendrick Coleman
    - http://www.amazon.co.uk/gp/product/B00BR07EBK
  - Storage Implementation in vSphere 5.0; Mostafa Khalil
    - http://www.amazon.co.uk/gp/product/B0091I7H1M
  - Essential Virtual SAN (VSAN): Administrator's Guide to VMware Virtual SAN; Cormac Hogan & Duncan Epping
    - http://www.amazon.co.uk/gp/product/B00LODTZSA

REMEMBER:

Storage is Boring

(but it's rather important!)

Alex Galbraith
@alexgalbraith
www.tekhead.org

TEKHEAD
IT'S ALL ABOUT THE WAF