Decoupling Storage From Compute in a Big Data Environment

Premise:

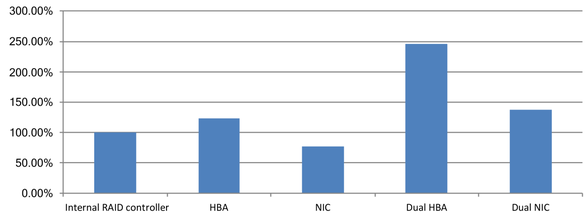
Big data solutions often rely on storage distributed inside of each individual compute node of the cluster. While cheaper per GB, this method forever ties the compute and storage together creating overspend if one of the pieces (either the compute or the storage) needs to be changed in scale. This is because the storage in a big data setup is most efficient when load is equally distributed avoiding the problem of certain servers needing to do more work. When compute and storage are tied together, a need to grow compute unnecessarily forces storage spend. Likewise storage increases force extra spend on computer hardware which can increase operational costs (more power) and support costs if the software is licensed per compute node.

SAN networks on the other hand are very expensive in these types of setups due to the volume of SAN switch ports and HBAs needed. A high amount of servers with distributed storage connected to a premium tier SAN is much more expensive than one server needing a SAN connection. This high cost is even more difficult to justify when the SAN provides no extra form of redundancy or survivability since often the software used to distribute the data across multiple compute nodes already includes measures to care for data redundancy and resiliency.

Our solution is to use a SAN network that is built upon IP instead of the traditional SAN protocols such as Fiber Channel (FC) and thus create a tier between the two extremes where the price delta is so small that decoupling the storage from the compute makes financial sense. There are savings recognized at every cost point (server adapter, switch port cost, ongoing support costs) and the solution has much greater scalability. Further refinement of the design of the network and even details such as cabling choice can greatly increase the savings. We have built a network with all the advantages of a SAN in terms of decoupling the storage from the compute while keeping costs to less than 2x the cost of traditional internal storage. (Traditional SAN costs, by comparison, are over 3x the cost of traditional internal storage).

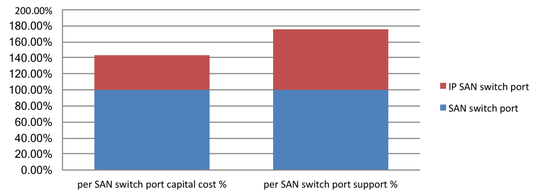
The example included is a 12 node compute cluster with 21.6 TB of usable storage across the 12 nodes (1.8 TB per node). A base line of 12 300GB 15k hard drives configured in a RAID 10 volume was considered the base cost of 100%. This number of nodes was also the base line number we used for our first round of throughput testing before doing other tests with an increased amount of nodes (see **Performance**). Note that costs are reflected as percentages and not actual dollars since every organization tracks costs differently including support costs, power consumption, rack space consumed, network support costs and other various factors.

Internal RAID controller vs storage adapter cost %



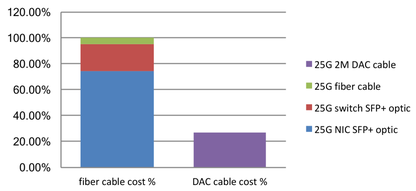
HBAs were 20% more expensive than an internal RAID controller. 25G NICs were 25% lower than an internal RAID controller. We went with redundant NICs since our use case was for production and not solely for research purposes.

Per SAN switch port capital cost and support %



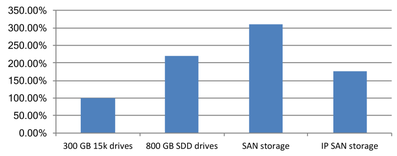
25G Ethernet ports were less than 45% of a 4 GB FC SAN switch port in terms of capital cost. For support costs, 25G Ethernet ports were 25% cheaper than a 4 GB FC port.

Fiber cable vs DAC cable cost %



Cost break down of one fiber optic 25G connection (host optic, switch optic and cable cost for 2Ms). One 2M Direct Attach Copper (DAC) cable was almost 75% cheaper. A DAC cable is a twinaxial or "Twinax" cable where the SFP+ adapters are directly attached to the Twinax with no optical conversion required.

Storage cost per GB %



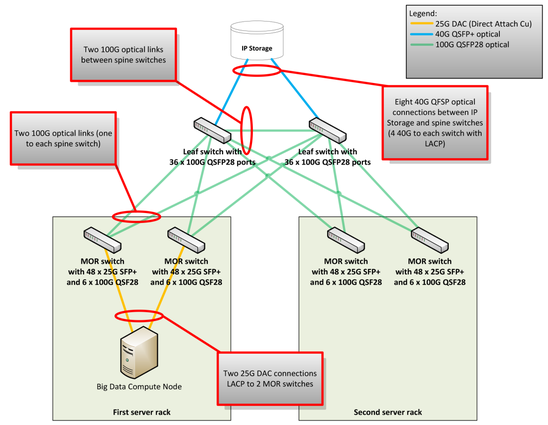
SAN storage was over 300% more cost when compared to internal disk. A flash-based IP SAN was 76% higher than internal disk. Using internal SSD flash-based drives was 120% higher than spinning media and 25% higher than a flash-based IP SAN solution.

Rack design

RMU	Description
48	Top Angle
47	
46	
35	
34	
33	[future 2U server]
32	
31	[future 2U server]
30	
29	[future 2U server]
28	
27	2U server
26	
25	1U leaf switch
24	1U leaf switch
23	2U server
22	
21	2U server
20	
19	[future 2U server]
18	
17	[future 2U server]
16	
15	
14	
2	
1	

Instead of a top-of-rack, or "TOR" data Ethernet switching, our design calls for data Ethernet switching to be mounted in the middle of the rack. We refer to this as "MOR" for middle-of-rack. Since DAC cable length causes a large jump in cost when using a longer length, a MOR solution allows for keeping every DAC cable in a rack to 2m. Servers can be installed above and below the center of the rack thus preventing the servers at the bottom third of a rack from needing to use longer DAC cables. For DAC cable cost comparisons, we found a 5m DAC cable costs 2x as much as a 2m DAC cable. This is very different from CAT6 copper wiring or fiber optic cables where the majority of the cost is in the connectors on the ends of the cable and not the length of the cable itself. DAC by contrast (at least currently) varies by a much higher ratio based on length compared to CAT6 or fiber cabling.

Logical Network Design



Middle-of-rack (MOR) leaf switches are fully meshed via 100G uplinks to both spine switches. Uplinks use Ethernet VPN (EVPN) which encapsulates the Layer 2 traffic inside of a Layer 3 routed connection so that both uplinks can be used to load-balance traffic.

Performance:

# of nodes	# of NFS mounts	IO size	sustained bandwidth	sustained IOPS	latency in ms
12	1	128 KB	9 GB/s	74,000	0.47
12	2	128 KB	15 GB/s	128,000	0.60
12	2	64 KB	12 GB/s	200,000	0.47
12	2	512 KB	16 GB/s	33,000	0.71
24	2	128 KB	24 GB/s	180,000	0.75

Conclusion:

Cost savings were recognized in the following areas:

- Use IP Ethernet SAN instead of FC
  - Cheaper storage adapter cost
  - Cheaper data switch port cost
  - Cheaper data switch port annual support cost
- Use DAC cables instead of fiber cables for 25G Ethernet
  - Fiber cost of server optic, switch optic and a 2m fiber cable is almost 4x as expensive as 2m DAC cables
  - Install switches as MORs instead of TORs to avoid needing longer DAC cables

Note that this solution does have a higher entry price just like buying a new SAN storage frame has a higher up front cost. We planned on this solution being extensible to many more compute nodes beyond the initial 12 which is why we focused on per server costs and did not hang the entire build cost of the network and storage device on just 12 servers. Some costs for items such as leaf switches were calculated based on total cost per leaf switch (capital for hardware, license and uplink connection costs) in order to accurately reflect the actual cost and not just the hardware capital cost of a switch port.

Carnegie Mellon University
Software Engineering Institute

Contact: **Tony Bichanich** **tony.bichanich@tdstelecom.com**
Matt Gruett **matt.gruett@tdstelecom.com**

Can IDS find the bad guys?

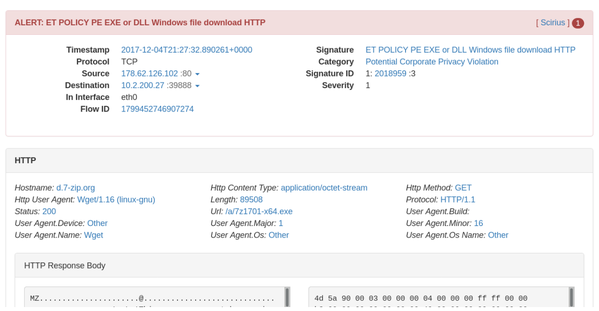
Yes, but they need to evolve.

Network Intrusion Detection System (IDS) like Suricata raises alerts containing information on the network flow triggering the alert but they don't provide an automatic way to find the source and target of attack. Signatures need to contain the information to allow better analysis.

Lateral Movement and Advanced Visualization. Taking over a network is usually done via a series of attacks. There is lateral movement when a source attacks a target and take control of it and use it as a source for other attacks. Reliable identification of source and target is a key toward advanced visualization and automatic discovery of attacks path.

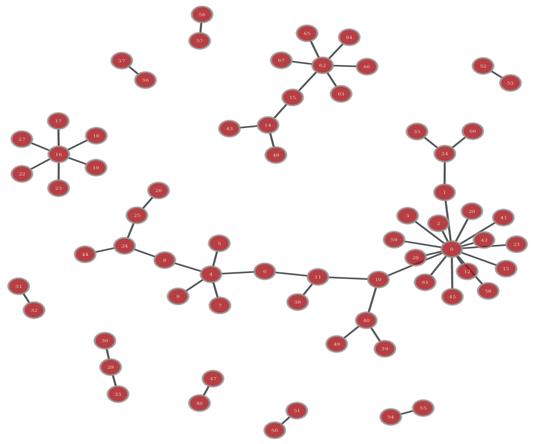
Find Source and Target of an Attack. The usual answer is the source is the source IP and the target is the destination IP. If this is true in most cases, this is wrong as soon as the signature matches on the answer to an attack. In this case this needs to be reverted.

Missing Information in Signature Language. The IDS signature language must evolve to contain indication of the source and target of attacks.



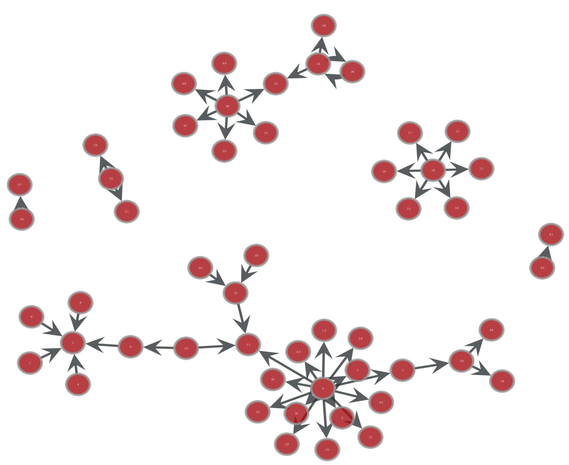
Alert event generated by Suricata in the EVE JSON format.

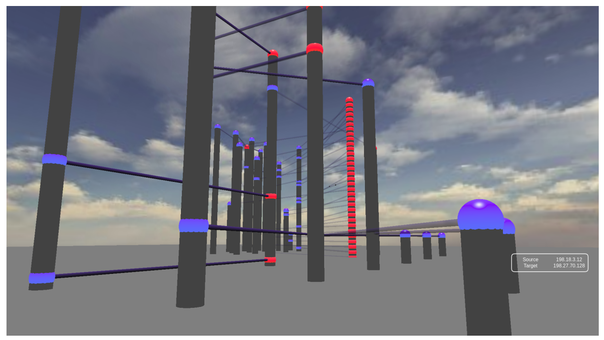
Building Non-oriented Graphs. We can build a graph where IP addresses are vectors and edge exist between two vectors if there is an alert where they are source and destination. Result on our data set shows really long possible paths.



Switching to Oriented Graphs. If we can get access to correct information about source and target we can create oriented graphs using the same idea but with edge going from source to target. The length of the possible paths is really shortened.

Oriented Graph Showing Paths





3D visualization using sources and targets with pktcity.js.

This representation displays all IP address as a cylinder. The z-axis is the time with latest time being the floor and past being high altitude.

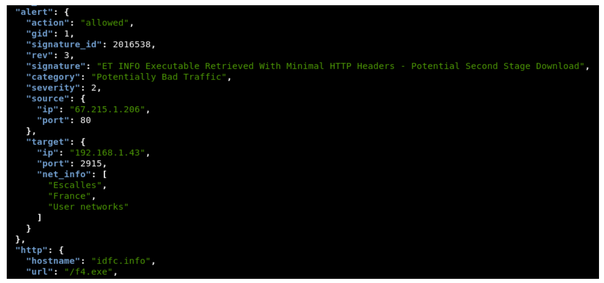
- Horizontal tunnels are alerts
- Ball travel in tunnel from source to target
- A red sphere means IP is source of an attack
- A blue sphere means IP is a target

An IP address appears on the graph as soon as it is source or target of an alert. The top of the cylinder is the timestamp of the first seen alert.

If an IP address as both blue and red colors then it is possibly a pivot point used for lateral movement.

Target keyword has been implemented in Suricata, the open-source threat engine developed by the OISF foundation a 501(c)3 non-profit foundation organized to build community and to support open-source security technologies like Suricata.

Introducing the target keyword. This new keyword, available in Suricata 4.0 and later in Snort++, allows the rules writer to define which side is the target of the attack. It can be set to either source or destination IP



Alert events now contain explicit source and target. When a signature contains the target keyword, Suricata adds a source and target sub-object in the JSON event that can be used by tools in their algorithms – critical to show organizational information allowing analysts to do statistics and analysis on the functional level.

Introducing a reliable way of knowing the source and target of attacks detected by an IDS is a great step toward better automated analysis of lateral movement and better visualizations.

Carnegie Mellon University
Software Engineering Institute

Contact: Éric Leblond el@stamus-networks.com