Administration & Operations

Hardware considerations

AEROSPIKE

Objectives

In this module we will cover

- Hardware considerations
- Aerospike Certification Tool (ACT)

Hardware - CPU

Aerospike is **multi-core** and **multi-threaded** and can take advantage of all available cores.

- Minimum: quad-core CPUs.
  - Common: 6 or 8 core CPUs.
- CPU is usually **not a bottleneck**.

top - if top shows software interrupts (**si**) **>30%**,

- indicates bound by a single network queue.

Hardware RAM

RAM is used for:

- Storage of primary index
- Storage of secondary index (if using)
- Storage of data (if configured RAM storage)

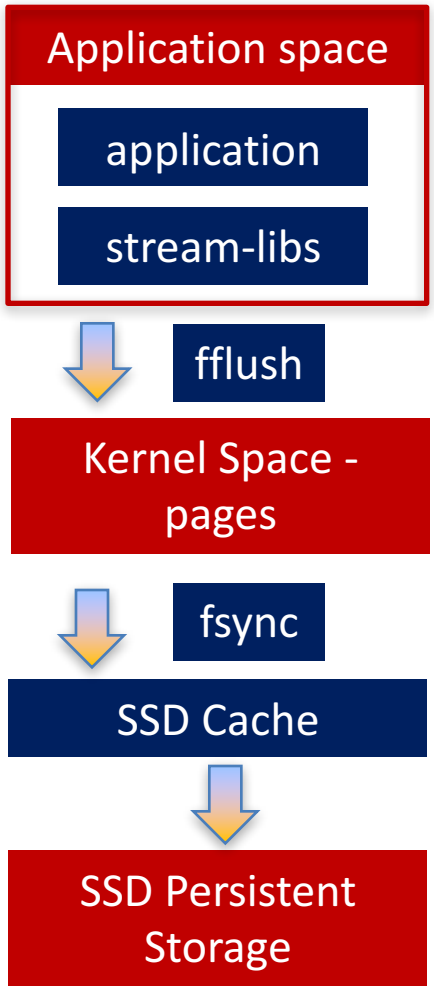
RAM speed **does not** significantly affect performance, unless TPS is **>~1Million** per node.

Hardware - Network

- **2 NICs**
  - 1 that is forward facing and is used to communicate with the clients.
  - 2 intra-cluster communication.
- Use **multicast**, if supported in your network.
- Do not underestimate the **bandwidth** need.
  - If you have 4 KB objects and need 30 ktps of transactions from a single node, this is 120 MB per second, or 0.96 Gb per second. This will obviously strain a 1 Gbps ethernet connection.
- **Link aggregation** for > 6Gbps bandwidth
  - Make sure the bonding method actually improves throughput, and is not a simple failover.
- **Irqbalance** – balance intrupts across multiple cores. (Linux kernels > 3.10)
  - Otherwise there are manual methods for doing the same thing. Contact Aerospike, if you wish to do this. There are hardware dependencies.

Hardware - RAID

- **Direct attach** of flash/SSD to the motherboard using SATA controller.
- Aerospike uses SSD as a block device – don't use a filesystem or format your SSD!
- If using RAID controller, use pass thru mode.
  - If you can't do pass thru and have to enable RAID 0 – setup each SSD as its own RAID 0. We don't want RAID controller doing striping (splitting data on multiple drives).
  - Disable SSD caching – we want to go from application space to persistent storage as fast as possible. Use write-thru with no read ahead.
- RAID controllers using the **LSI 2208** chip can take advantage of StorCLI or MegaCLI (**LSI Fastpath)**.
- PCIe based flash devices are still new – very fast.
- Overprovision your SSD device (21% recommended)
- Partition and initialize



http://www.aerospike.com/docs/operations/plan/ssd/ssd_setup.html

Testing SSDs with the ACT

SSD Testing

Aerospike has specialized the database to make the most out of SSDs. But not all SSDs have the same performance. Getting the most means being able to test.

Aerospike has open sourced the Aerospike Certification Tool (ACT), which is available at Github (https://github.com/aerospike/act).

What the ACT Does

The ACT is a low level tool used to test the performance of SSDs as Aerospike uses them. The default settings for the tests are:

- **Reads** are of 1.5 KB objects
- **Writes** are in large blocks of 128 KB
- **Simultaneous** Reads and writes
- Simulates Aerospike **garbage collection**
- Does not use network
- Published results are for a **single drive**.
  - Aerospike can use these in parallel, but linearity depends on several factors in the hardware.
- Tests will work on a factor of "x". A good SATA drive in 2015 will support 3x speeds.
  - "1x" was the performance of a good SSD in 2011. This is the equivalent of 2,000 reads/second with 1,000 simultaneous writes/second.
  - "2x" is twice that, "3x" is 3 times, etc.

What Affects ACT Tests

- SSDs are not all alike.
  - Higher model numbers are not always better.
- Changing the object size, read/write ratio, etc… will change the results.
- RAID controller – How the drives are connected is extremely important.
  - Low cost controllers will **always add ms of latency**.
  - Even high cost RAID controllers may have issues. If you are using one of these, please see our note at: http://www.aerospike.com/docs/operations/plan/ssd/lsi_megacli.html
- Performance can depend a lot on **overprovisioning** (OP). Total OP should be between 22%-29%.
- Generally, **newer firmware** is better.

Running ACT Tests - Preparation

- ACT will **destroy data** on the drive/partition.
  - Make sure you are not using the drive with the OS or any important data.
- Use the **correct device** IDs
  - e.g. /dev/sdb, /dev/xvdb, etc.
- Tested devices at http://www.aerospike.com/docs/operations/plan/ssd/ssd_certification.html
- Download and compile the ACT (https://github.com/aerospike/act).
- Before running any tests, run the program **`actprep`** on the device/partition under test. This will place random data on the device.
- If necessary OP the drives (instructions are at: http://www.aerospike.com/docs/operations/plan/ssd/ssd_setup.html).

Overprovisioning (OP)

Overprovisioning is space on an SSD that is reserved for use by the SSD controller (not the RAID controller).

Total OP (manufacturer + user) for Aerospike use should be at least 29%.

- Consumer drives – generally have 6%-8%
- Enterprise SATA – vary a lot, but many have up to 30%
- PCIe/NVMe – 22%+

This space is not user accessible and user OP will **decrease the size** of the SSD.

Running ACT Tests – Executing Tests

- Create a configuration file for the test by executing:

```
[act-3.0]$ python act_config_helper.py
Enter the number of devices you want to create config for: 1
Enter either raw device if over-provisioned using hdparm or partition if over-provisioned using
fdisk:
Enter device name #1 (e.g. /dev/sdb or /dev/sdb1): /dev/sdb
Change test duration default of 24 hours? (y/N): y
Enter the test duration in hours: 3
Use non-standard configuration? (y/N): n
"1x" load is  2000 reads per second and 1000 writes per second.
Enter the load factor (e.g. enter 1 for 1x test): 3
Do you want to save this config to a file? (y/N): y
Config file actconfig_3x_1d.txt successfully created.
```

- Select an appropriate starting point for your drive
  - Consumer SATA/Cloud – 3x
  - Enterprise SATA – 6x
  - PCIe/NVMe – 24x
- Run the ACT test and redirect output to a file:

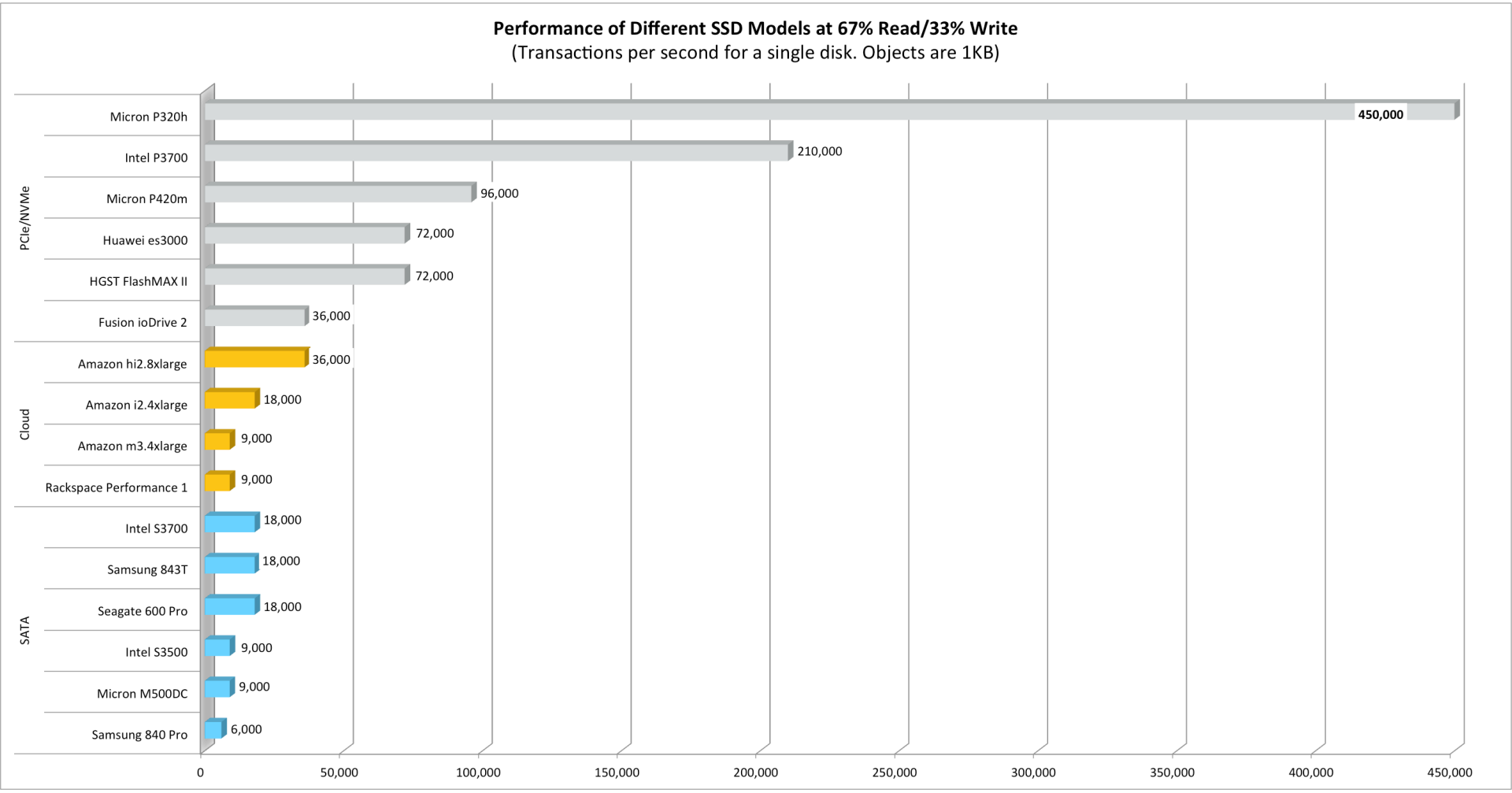
```
[act-3.0]$ act actconfig_3x_1d.txt > actconfig_3x_1d.log &
```

- Change the x factor and retest. The final test should be for 24 hours.

Analyzing the results:

- Pass is:
  - < 5% of transactions should take longer than 1 ms
  - < 1% of transactions should take longer than 8 ms
  - < 0.1% of transactions should take longer than 64 ms

SSD Certification Tests



Aerospike maintains a list of SSD performance numbers at:

http://www.aerospike.com/docs/operations/plan/ssd/ssd_certification.html

Summary

What we have covered:

- Hardware considerations
- Aerospike Certification Tool (ACT)