Topics

- Tuning and Workflows
- Networking
- Storage
- Architecture
Tuning and DTN Workflows

• Why do we need to tune
• Requirements depend on the workflow
• Determine what you are trying to optimize
  • User experience
  • Resource utilization
  • Efficiency
• User Experience as Sender / Receiver
• Simultaneous Users
• Fairness between DTNs
Tuning

• Test before and after tuning
• Tuning for production vs testing
• Tools
  • `sysctl` and `proc`
  • `iproute2`
  • `ethtool`
  • BIOS settings
  • Vendor specific tools (e.g. `mlxlink`)
Tuning Kernel Parameters

• **Sysctl**
  • sysctl command
  • /etc/sysctl.conf and /etc/sysctl.d/

• proc filesystem
  • /proc and /proc/sys

# sysctl net.ipv4.tcp_available_congestion_control
net.ipv4.tcp_available_congestion_control = reno cubic

$ cat /proc/sys/net/ipv4/tcp_available_congestion_control
reno cubic

$ cat /proc/sys/net/ipv4/conf/eth0.99/forwarding
# sysctl net.ipv4.conf.eth0.99.forwarding
Networking

- Network parameters on the host
- Socket Parameters
- Protocol Specific Parameters
- Driver Settings
- MTU
  - Detection
  - Jumbo frames
- Traffic Control (tc)
  - Queue Discipline
### Socket Parameters

- Apply to all protocols
- Set with `sysctl` or proc filesystem
- Send and Receive Buffer sizes
  - `net.core.rmem_default`
  - `net.core.wmem_default`
  - `net.core.rmem_max`
  - `Net.core.wmem_max`
- Privileged applications can ignore limits
TCP Parameters

• Set with sysctl or proc filesystem
• Send and Receive Buffer sizes
  • net.ipv4.tcp_rmem
  • net.ipv4.tcp_wmem
  • Three values: minimum, default, maximum
  • Default value overrides net.core.[r|w]mem_default
• TCP Memory Management
  • net.ipv4.tcp_mem
  • Three values: low, pressure, high
TCP Parameters

• Congestion Control
  • net.ipv4.tcp_available_congestion_control
  • net.ipv4.tcp_allowed_congestion_control
  • net.ipv4.tcp_congestion_control
  • Kernel module may need to be loaded
  • Privileged program may use any available
  • ss utility provides information on existing TCP connections
Network Driver Setting

- Send and Receive Ring Buffers
- Offloading features
- ethtool
- Vendor tools (mlxlink)

```
# ethtool --show-ring enp33s0f0
Ring parameters for enp33s0f0:
Pre-set maximums:
RX: 8192
RX Mini: 0
RX Jumbo: 0
TX: 8192
Current hardware settings:
RX: 1024
RX Mini: 0
RX Jumbo: 0
TX: 1024
```

```
# ethtool --show-features enp33s0f0
Features for enp33s0f0:
rx-checksumming: on
tx-checksumming: on
tx-checksum-ipv4: on
tx-checksum-ip-generic: off [fixed]
tx-checksum-ipv6: on
tx-checksum-fcoe-crc: off [fixed]
tx-checksum-sctp: off [fixed]
scatter-gather: on
tx-scatter-gather: on
tx-scatter-gather-fraglist: off [fixed]
tcp-segmentation-offload: on
tx-tcp-segmentation: on
tx-tcp-ecn-segmentation: off [fixed]
tx-tcp-mangleid-segmentation: off
tx-tcp6-segmentation: on
udp-fragmentation-offload: off
generic-segmentation-offload: on
generic-receive-offload: on
large-receive-offload: off
```
Storage

• Types
  • Spinning Disks
    • Larger Capacities
  • Solid State Drives
    • Faster Internally
  • Differences in durability

• Interface
  • SATA - 6 Gb/s
  • SAS - 12 Gb/s
  • NVMe - 32 Gb/s (PCIe 3.0 4x) (SSD only)
  • SATA disk compatible with SAS interface
Storage Speeds

• Internal drive speeds typically slower than interface speed
  • Drive buffers used to compensate
• Accessing data sequentially provides better internal throughput
  • Some vendors provide specs for sequential and random throughput
  • SSD does not necessarily need to be sequential (spatial locality)
• Read performance is typically better than write
  • Read/Write optimized drives
• Performance may degrade over time
  • Thermal throttling
  • TRIM
Storage Tuning

- Physical devices can be logically grouped to increase performance
  - RAID 0
  - Software RAID

- Memory can be used to improve apparent storage performance
  - Short term
  - Read Caches
  - Write Buffers

- Performance differs depending on the file system in use
  - Large files versus lots of small files

- TRIM/discard implementation
  - OS, files system, and device dependent

- Sector size
  - 512e vs 4Kn
Architecture

• Fast vs Many Cores
  • Depends on workflow
  • Single/Multi-Threaded application
    • Multi streams is not necessarily multi threads
  • Simultaneous users

• PCI Express
  • Lane availability (Intel vs AMD)
  • PCIe 4.0 availability and support
  • PCIe switches
Architecture

• Non-Uniform Memory Access (NUMA)
  • Physical location of devices within a system matters
  • numactl and numad utilities

• Power Saving
  • Can be disabled to improve performance
  • Different setting for production and testing
  • cpupower utility

• IRQs
  • Handle close to the hardware (NUMA node)
  • Avoid bottlenecks
  • irqbalance utility
Thank you

Any questions?

www.geant.org