ISUG-TECH 2015 Conference

SAP IQ Hardware Sizing and Internals

Mark Mumy, SAP

Agenda

Agenda

- Welcome
- Speaker Introduction
- Memory
- CPU
- Disk
- Network
- Q&A

Who is Mark Mumy?

- Came to SAP via the Sybase acquisition
- 19+ years with SAP
- Over 19 years experience with SAP IQ, 4+ years with HANA, 3+ years with Big Data
- Dabbled in replication, streaming, business intelligence, ETL
- Been involved in at least half of the SAP IQ architectures
- Author of numerous whitepapers and many TechWave, TechEd, d-code, ISUG Tech sessions
- Spent most of my time focusing on EDW, data marts, ODS, Big Data, high speed/high throughput database computing
- Chief architect on the Sybase/SAP IQ Guinness World Record systems
- Don't hesitate to contact me!   Email me at ***mark.mumy@sap.com***

TERMINOLOGY

- I tend to use the term core, processor and CPU interchangeably. What I mean is a physical processor core that does work. Not a thread, not a socket, not a processor.
- Physical Volume – If a volume manager is being used, this is the LUN or storage unit that is used to build the volumes
- Logical Volume – The logical end result of a volume manager and the LUNs (physical volumes) that it manages
- Volume Group – A set of one or more physical volumes from which space can be allocated to one or more logical volumes

QUICK SIZING REFERENCE

- **<u>RAM</u>**: 8-16 GB per core (8-12 for simplex, 12-16 for MPX)
- **<u>RAM</u>**: Give IQ 85-90% of available RAM (assumes there are no other major consumers of RAM on the host)
- **<u>Storage</u>**: Prefer RAID 10 for write intensive system and temp store
- **<u>MAIN Store disk</u>**: 2-3 drives per core on the host or in the **entire** multiplex. Assuming 75-100 MB/sec throughput per core.
- **<u>TEMP Store disk</u>**: 2-3 drives per core on the host. Assuming 75-100 MB/sec throughput per core.
- **<u>MAIN Store Fiber Controllers/HBAs</u>**: 1 per 5-10 cores. Size based on total MB/sec throughput needed on host.
- **<u>TEMP Store Fiber Controllers/HBAs</u>**: 1 per 5-10 cores. Size based on total MB/sec throughput needed on host.

Quick Sizing Reference

Memory

- ✓ 8-12 GB per core for simplex
- ✓ 12-16 GB per core for multiplex
- ✓ Main cache: 30% of total RAM
- ✓ Temp cache: 30% of total RAM
- ✓ Large Memory: 30% of total RAM
- ✓ RLV: allocate as necessary

Disk

- ✓ 50-100 MB/sec IO throughput needed per core
- ✓ Allocate HBAs as necessary (minimum of 2 for redundancy) for total throughput
- ✓ Could be direct attached, fiber channel, ethernet (see below), or proprietary

Network

- ✓ Minimum of 2 x 10 gbit ethernet (one public, one private)
- ✓ Add more, as necessary, if storage is over the network (NAS, NFS, FCoE, etc)

MEMORY

- Most important rule is to NOT OVERCONFIGURE MEMORY
  - Swapping leads only to horrible performance in IQ
- First, understand the "memory map" for your server
  - Operating System
  - OLAP Servers
  - Middleware
  - Other applications
  - Monitoring applications
  - Shells, scripts, etc
  - File system buffering
- Deduct all of this from the total memory, and go from there
- A good starting point if a memory assessment cannot be done is to configure the combined total of the IQ caches for no more than 66% of total RAM on the machine

MEMORY (CON'T)

- Sybase IQ memory will consist of the following:
  - Catalog cache (-c/-ch/-cl options in the configuration file)
  - Thread memory (stack size * number of IQ threads)
  - Main cache (-iqmc option in the configuration file)
  - Temporary cache (-iqtc option in the configuration file)
  - Version memory
  - Load memory (reduced in IQ 15.0 and eliminated in IQ 15.2)
  - Memory used during backups

MEMORY (CON'T)

- Catalog cache
  - Generally, set this to 2-8 times the size of the catalog file
    - Rule of thumb default is 64-128 MB for most systems
    - Set higher for highly concurrent systems
  - TLV replay uses catalog cache
    - Performance can suffer if catalog cache is too small

MEMORY (CON'T)

- This option was greatly reduced in use for IQ 15.0 and 15.1 and has been completely eliminated as of IQ 15.2
- Load_Memory_MB
  - Allowable values: 0 (unlimited) to 2000
  - Calculation is: TableWidth * 10,000 * 45 / 1024 / 1024
    - TableWidth is the binary width of the table as collected via *tablewidth()* function in IQ.
    - 45 – Total number of buckets to store data
    - 10,000 – Default number of rows for each bucket
  - General recommendation is to not let this value default to the above formula
    - May hinder some loads or be over configured for other loads
    - The positive impact to the OS should outweigh the possible side effects
    - Presents a consistent memory map to the OS which greatly reduces OS memory fragmentation
    - OS level memory fragmentation will cause IQ to appear to have a memory leak in that IQs memory footprint will grow even though sp_iqstatus shows a much lower value

MEMORY (CON'T)

- Cache Memory Used During Loads
  - Memory allocation from the main cache
    - 1 page for each FP index + 1 page for each distinct value in LF indexes
  - Memory allocation from the temp cache
    - Only HG and WD indexes will use temp cache during loads
    - For HG: ***(8 + sizeof(datatype)) * numberRowsBeingLoaded***
    - WD indexes memory use will be substantially more because each word (token) in the data value requires some temporary cache
    - Each token would require the same memory as the HG index
    - ***~ numberTokens * (8 + sizeof(datatype)) * numberRows***

MEMORY (CON'T)

- Bitmap Memory
  - Additional heap memory allocated during loads for storing bitmaps – exclusive of LOAD_MEMORY_MB
  - Applies to LF, HNG, DTTM, DATE, TIME, and CMP indexes
  - Total amount of memory is dependent on number of distinct values, which is not known before load begins
    - Makes this memory consumption impossible to predict for IQ
  - Groups, or chunks of memory for these bitmaps are allocated in 8k increments
    - Using an example of 500 LF indexes and assuming N distinct values per column, the virtual bitmap memory consumed is:
      - 8,192 * 500 * N = 4,096,000 * N ~ 400MB!!!
      - Can be controlled slightly with LF_BITMAP_CACHE_KB (for LF)

MEMORY (CON'T)

- Version Memory
  - Dynamic RAM allocated and freed when needed
  - Generally, the amount is small
  - Can grow to significant levels (hundreds of MB to GB) when the server has a lot of old versions around
  - Readers can have version memory increases for work performed on writer(s)

MEMORY (CON'T)

- Backup Memory
  - In ideal situation, amount of memory used during a backup is a function of
    - number of cpus
    - number of main or local store dbspaces to be backed up
    - block factor
    - IQ block size (as seen in column 'block_size' in sys.sysiqinfo)
  - Approximate memory used by backup process (z) will be
    - **y = max(2 * number_of_cpus, 8 * number_of_main_or_local_dbspaces)**
    - **z = (y * 20) * (block factor * block_size)**

MEMORY (CON'T)

- Backup Memory (example)
  - dbspaces = 50
  - block factor = 100
  - number of cpus = 4
  - block_size = 8,192

  - 'y' is max(8, 400) ➔ y=400
  - 'z' is (400 * 20) * (100 * 8,192) ➔ 6.5GB

- This memory comes entirely from the OS and is not released until the entire backup operation completes
- Block factor setting is the primary way of controlling this, but of course there are performance tradeoffs when doing this

MEMORY (SUMMARY)

Operating System	**.5 to 1 GB RAM**
Filesystem Cache	**5-10% of RAM**
All Other Applications	
IQ Catalog Memory	**-c/-cl/-ch parameters**
IQ Thread Memory	**stack size * thread count**
Large Memory	**30% of remaining RAM**
Bitmap Memory	**per concurrent load**
IQ Main Cache	**30% of remaining RAM**
IQ Temporary Cache	**30% of remaining RAM**
Backup Memory	**per backup instance**

MEMORY - IQ 15 CHANGES

- Load Memory – Load memory is still employed in IQ 15. Most of the memory, though, comes from temporary cache. A significantly smaller amount now comes from heap space. This has been completely removed from the product as of version 15.2
- User Defined Functions – Any memory used by the UDF/UDAF framework is done outside the IQ caches. It is allocated when the UDF starts and is freed when the UDF exits.
- The new 3-byte FP indexes use more main cache during creation and data change. The 1-byte and 2-byte FP indexes also used memory, but it was generally insignificant. Due to the maximum cardinality (2^24) this can be a large amount of RAM.

MEMORY - IQ 16 CHANGES

- Load Memory – Completely removed
- All data loading was moved from temp cache (v15) into the new large memory accumulator (LMA, -iqlm)
- LMA also hosts the n-bit lookup table structures. In v15, the lookup tables for FPs were in main cache.

DATA LOAD CPU & MEMORY SIZING

- Single row inserts / updates / deletes
  - Use very little CPU / memory resources (not significant in the overall scope of discussion)
- Bulk load inserts / updates / deletes
  - Includes LOAD TABLE, INSERT…FROM LOCATION, INSERT SELECT, and UPDATE, and multi-row DELETES
  - For ***maximum*** performance (assumes 100% availability of the CPU resources and no contention)
    - 1 CPU for every 5-10 columns in the table being loaded (default FP index)
    - 1 CPU for every 5-10 indexes (HNG, LF, CMP, DATE, TIME, DTTM) on the table that have not been mentioned
    - HG, WD, and TEXT indexes can consume all cores on the host during pass 2 of the loads!
  - These "ideal" recommendations should be balanced against the load performance requirements at any individual site

DATA LOAD CPU & MEMORY SIZING

Version 15

- Alternative Algorithm (data volume based)
  - For systems with 4 or fewer CPUs, expect to load roughly 10 GB of data per hour per CPU
    - A 4 CPU system should be able to load about 40 GB of raw data per hour
  - For systems with 8 or more CPUs, expect a load rate of 20-50 GB per hour per CPU
    - An 8 CPU system should be able to load between 160 and 400 GB of raw data per hour
  - Load times with this approach will vary greatly based on CPU count / speed and the number and types of indexes on the table being loaded
  - For each BLOB or CLOB being loaded into IQ a single CPU will be necessary for maximum performance

DATA LOAD CPU & MEMORY SIZING

Version 16

- Forget the previous rules!
- 10-20 MB of raw, file based data loaded per second, per core on the host
- A 40 core host should load 400-800 MB of data per second

DATA LOAD CPU & MEMORY SIZING

- Memory – Temp Cache Load Requirements
  - As always, more is generally better!
  - During loads, HG indexes will use temporary cache to store the intermediate data necessary for the HG indexes
  - During pass one of the load process, the data necessary to build the HG index is stored in the temporary cache
  - Should there not be enough temporary cache to store the data, the server will flush pages to disk
  - For all columns that contain an HG or WD index, temp cache memory requirement is roughly

total_pages = 1 + ((number_of_rows_in_load_files * width_of_columns_in_hg) / page_size_in_bytes)

DATA LOAD CPU & MEMORY SIZING

- Memory – Temp Cache Load Requirements (con't)
  - As an example, let's assume that a load of 10 million rows is taking place on a table that contains a single HG index on an integer column and that the database was created with a 256K page
  - The total temporary cache necessary for this load would be:
    - total_pages = 1 + ((10,000,000 * 4) / 262,144)
    - total_pages = 1 + (40,000,000 / 262,144)
    - total_pages = 153 or 38 MB

DATA LOAD CPU & MEMORY SIZING

- Memory – Main Cache Load Requirements
  - FP ➔1 page for each FP index (plus 3 in temp cache for each optimized FP)
  - LF ➔ 1 page for each distinct value currently being loaded into the LF index
  - HNG, DATE, TIME, and DTTM ➔ 1 page for each bit in the bitmap
  - CMP ➔ 3 per index
  - HG and WD ➔ Reliant on temporary cache during the first pass (see above) and the main cache during the second pass to build the final page structures
    - There is minimal main cache needed for HG and WD indexes due to the reliance on temporary cache

DATA LOAD CPU & MEMORY SIZING

- Memory – Main Cache Load Requirements
  - A rough rule of thumb is 5-10 pages of main cache need to be allocated for each index on the table being loaded (this includes all index types, including the default FP index)
  - Example: For a table with 50 columns and 25 additional indexes, this would equate to:
    - (50 FP indexes + 25 additional indexes) * (5,10) → 375-750 pages
    - 375 pages * 128K page size → 46.875 MB
    - 750 pages * 128K page size → 93.75 MB

DATA LOAD CPU & MEMORY SIZING

The net of it all for IQ 16?

- Give 33% of the RAM to main cache
- Give 33% of the RAM to temp cache
- Give 33% of the RAM to LMA cache

- That is 33% of the RAM left over.
- For a system dedicated to IQ with little other activity, I assume that 10% of the RAM will be used by the OS. 1/3rd of the remaining 90% would be 30% of the machines RAM.
- For a 100GB RAM system, that's 30gb for each of the caches above and 10gb for everything else.

QUERY SYSTEM CPU & MEMORY SIZING

- IQ 15 changed the game with respect to query operations
- IQ 16 is continuing to change and expanding with every release
- Most queries are now done in parallel consuming all resources on the host
- In general, IQ will try to blend single and multi-user query performance
- As more queries appear on the run queue, the available resources will change and IQ will adjust accordingly

QUERY SYSTEM CPU & MEMORY SIZING

On to the old way of sizing queries…

- It's important to 'classify' queries in order to more accurately size the hardware
  - Super fast – Queries that generally take less than five seconds to run
  - Fast – Queries that generally take less than three minutes to run
  - Medium – Queries that generally take between three and 10 minutes to run
  - Long – Queries that generally take more than 10 minutes to run
  - Super long – Queries that generally take more than 30 minutes to run
- General rule of thumb: Most queries will consume between 1 and 2 CPUs for the duration of their execution

QUERY SYSTEM CPU & MEMORY SIZING

- Now that the types of queries have been defined, we need to apply the types of queries to CPUs on the host.
  - Super fast – Each CPU should be able to handle 10 or more concurrent queries
  - Fast – Each CPU should be able to handle between five and 10 concurrent queries
  - Medium – Each CPU should be able to handle between two and five concurrent queries
  - Long – Each CPU should be able to handle between one and two concurrent queries
  - Super Long – Each CPU should be able to handle at most one query, generally a query will need more than one CPU

QUERY SYSTEM CPU & MEMORY SIZING

- One caveat is parallel execution
  - Some simple queries will be broken up automatically and run over multiple CPUs assuming appropriate resources are available
- As more queries appear on the run queue, the available resources will change and IQ will adjust accordingly
- In general, IQ will try to blend single and multi-user query performance
- When sizing for query performance, a minimum of 4GB RAM per CPU is recommended
  - For smaller systems with less than 8 CPUs, 4-8GB should be the recommendation
- Of course, multiplex adds a new dimension to this discussion, but the general rules of thumb still apply!

DISK & I/O SIZING

- Disk sizing requirements change as CPU speeds change
  - These guidelines take a "middle of the road" approach
  - If you're sizing for a faster or slower platform, take that into consideration
- Remember, IQ tends to be CPU-bound, rather than I/O bound
  - As CPU speeds and throughput increase, it drives the bottleneck closer to the disk subsystem
- Disk strip size should be 64k or larger
  - In general, use the largest stripe size available to the disk subsystem, particularly when using larger (256k or 512k) IQ page sizes in conjunction with larger (multi-terabyte) databases
- Try to avoid multiple FC HBA's utilizing the same system bus connector

DISK & I/O SIZING

v15

- On average, a typical CPU can ingest 20MB of data per second from Sybase IQ
  - As a rule of thumb, design the disk farm to deliver 20MB/sec to all CPUs in the ***entire multiplex environment***
- No LVM should be used if possible!
  - No benefit for IQ, decreases performance, and can increase cost!
- Use larger disks (146GB, 300GB, 500GB, 750GB)
  - Lower RPM (7200RPM), larger drives have been tested but can slow down write intensive systems

DISK & I/O SIZING

v16

- On average, a typical CPU can ingest 20-200MB of data per second from Sybase IQ
- Not all systems need this much bandwidth
- Current sizing guidance is to guarantee 50-100 MB/sec of IO per core in the entire multiplex

DISK & I/O SIZING

- When using Fiber Channel Drives
  - Typical Use Environment: 0.3 – 0.5 spindles for every CPU in the multiplex environment
  - Heavy Use Environment: 1-2 spindles for every CPU in the multiplex environment
- For SATA Drives
  - Typical Use Environment: Minimum 1 spindle for every CPU in the multiplex environment
  - Heavy Use Environment: 2-4 spindles for every CPU in the multiplex environment
- For SAS Drives
  - Typical Use Environment: Minimum 0.5 spindle for every CPU in the multiplex environment
  - Heavy Use Environment: 1-3 spindles for every CPU in the multiplex environment

DISK & I/O SIZING

v16

- Drive types don't matter
- They do matter in terms of performance, but don't matter to IQ
- The net for v16 is that we want 50-100 MB/sec throughput per core
  - That could be 1 SSD per core, 2-3 fiber channel drives, or 2-5 SATA drives

DISK & I/O SIZING

v15

- Disk Controllers
  - 1 per 5-10 CPU's for typical use, more for heavy use environment
- Don't forget to make sure the disk subsystem can support the overall SAN to bandwidth requirements as well

  number_of_total_cores * 20 MB/sec

- **None of the disk sizing takes in to account the overhead for RAID levels**
  - RAID 5 will require 1 additional drive per RAID 5 group (parity drive)
  - RAID 0+1/1+0 will require twice as many drives (mirrors)

DISK & I/O SIZING

v16

- Disk Controllers
  - 1 per 5-10 CPU's for typical use, more for heavy use environment
- Don't forget to make sure the disk subsystem can support the overall SAN to bandwidth requirements as well

  number_of_total_cores * 50-100 MB/sec

- **None of the disk sizing takes in to account the overhead for RAID levels**
  - RAID 5 will require 1 additional drive per RAID 5 group (parity drive)
  - RAID 0+1/1+0 will require twice as many drives (mirrors)

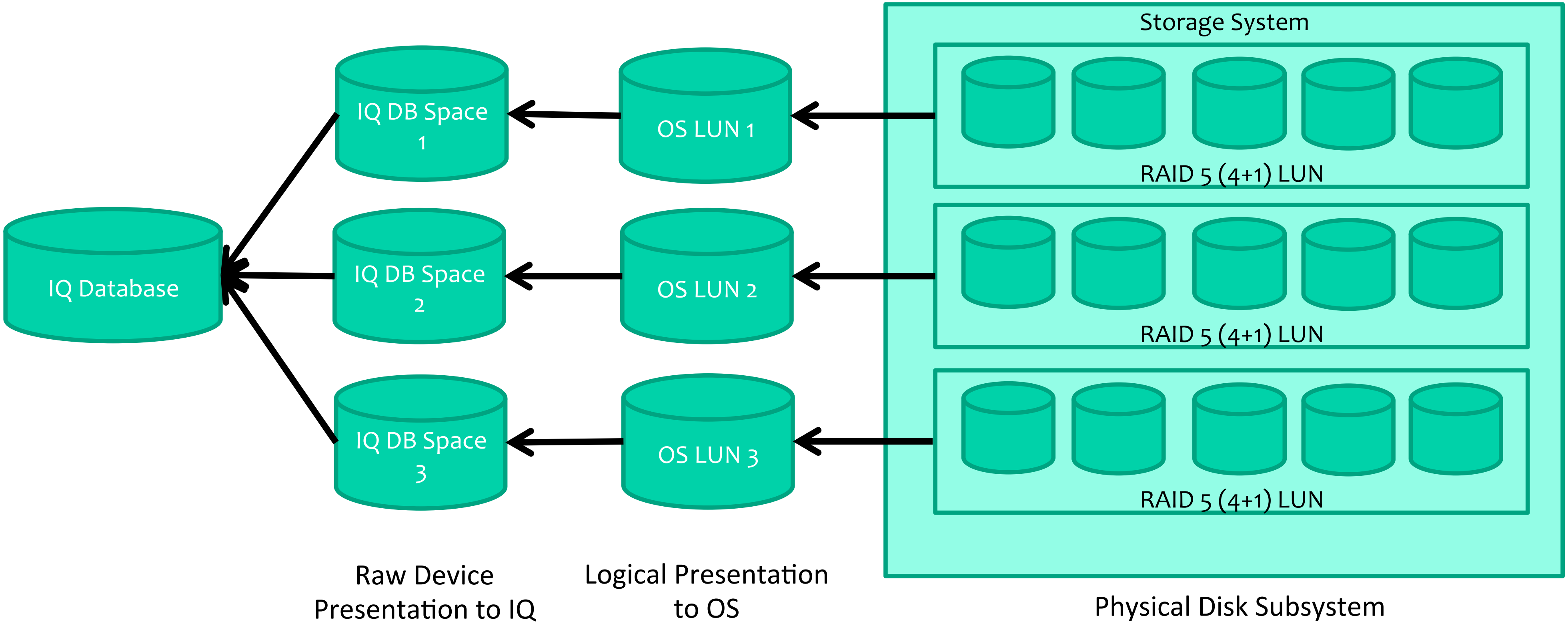
DBSPACE CHANGES – IQ 15 AND LATER

- A dbfile in IQ 15 is synonymous with an IQ 12.x dbspace
- A dbspace in IQ 15 is a logical grouping of storage (dbfiles)
- IQ_SYSTEM_MAIN still exists
  - Do not place user data on this dbspace
    - Create a user defined dbspace (1 or more) for all user data
  - Shared area for
    - Free list for all dbfiles in all dbspaces
    - Versioning, limited node-to-node communication, TLV replay, DDL
    - By default, 20% of this space is reserved for freelist and TLV
  - Adding space to IQ_SYSTEM_MAIN currently forces the entire MPX to be shutdown and synchronized

DBFILE CONSIDERATIONS

- **Do not confuse our dbspace requirements with the physical drive requirements! They are different and should be treated separately.**
- For Main Store:
  - **IQ_dbfiles = 8-12 dbfiles, minimum**
- For Temp Store:
  - **IQ_dbspaces = 4 dbfiles, minimum**
- Always try to physically separate the spindles used for Main and Temp store devices
  - This allows more fine-grain tuning at the device level based on the very-different read/write characteristics of Main and Temp store devices (assuming the SAN has this capability)

DRIVE MAPPING PICTURE



NETWORK CONSIDERATIONS

- Not of huge importance, but can be significant with respect to
  - INSERT…LOCATION performance
  - Large result data sets returning across the network
- Consider that movement of 100MB of data will require:
  - 80 seconds on a 10 Megabit LAN
  - 8 seconds on a 100 Megabit LAN
  - 0.8 seconds on a Gigabit LAN
- Consider using faster network cards, better network topology, dedicated switch / hub for servers, etc…
- Increase the packet size of the client application (the method varies depending on ODBC, JDBC, or Open Client connectivity)
- Increase the packet size of the remote data load by using the 'packetsize' parameter to the insert…location syntax

THREADS

- Total IQ threads allocated at startup is based on
  - Number of connections (-gm)
  - Number of cores (-iqnumbercpus)
- By default, -iqmt is set to:

  60*(min(numCores,4)) + 50*(numCores - 4) + (numConnections + 2) + 6

THREADS (CON'T)

- Two main types of IQ threads
  - Connection Threads
    - 2*(numConnections + 2)
    - Reserved for connections
  - Server Threads
    - 60*(min(numCores,4)) + 50*(numCores - 4)
    - Support load and query operations
- Total threads can be set via –iqmt
  - Make sure that –iqmt is larger than total threads needed for connections!
  - Upper limit is currently 4096

THREADS (CON'T)

- Threads to handle I/O
  - Pulled from the Server Thread pool
  - Two types of I/O threads
    - Sweeper – write dirty buffers to disk
    - Prefetch – read data from disk into cache
  - SWEEPER_THREADS_PERCENT – default is 10% of total threads (-iqmt)
  - PREFETCH_THREADS_PERCENT – default is 10% of total threads (-iqmt)

THREADS FOR I/O OPERATIONS

- When sweeper threads fall behind the wash area, dirty buffers are used
  - When dirty buffers are used, the query or load thread must now perform the disk write instead of the sweeper threads
- Prefetch threads can fall behind the prefetch requests
  - When this happens, the query or load must stop processing and get the buffer from disk directly
- In an ideal configuration, sweeper and prefetch threads would be the only threads doing disk I/O.

SUMMARY

What does this all mean?

- The formulas are just a starting point for good performance
- Every application and system is different so adjust accordingly
- Loads will consume enough cores to get the job done
  - HG/WD/TEXT indexes, though, can use all cores on the host
- Queries will generally consume all cores on the host

Quick Sizing Reference

Memory

- 8-12 GB per core for simplex
- 12-16 GB per core for multiplex
- Main cache: 30% of total RAM
- Temp cache: 30% of total RAM
- Large Memory: 30% of total RAM
- RLV: allocate as necessary

Disk

- 50-100 MB/sec IO throughput needed per core
- Allocate HBAs as necessary (minimum of 2 for redundancy) for total throughput
- Could be direct attached, fiber channel, ethernet (see below), or proprietary

Network

- Minimum of 2 x 10 gbit ethernet (one public, one private)
- Add more, as necessary, if storage is over the network (NAS, NFS, FCoE, etc)

QUICK SIZING DETAILED REFERENCE

- **RAM**: 8-16 GB per core (8-12 for simplex, 12-16 for MPX)
- **RAM**: Give IQ 85-90% of available RAM (assumes there are no other major consumers of RAM on the host)
- **Storage**: Prefer RAID 10 for write intensive system and temp store
- **MAIN Store disk**: 2-3 drives per core on the host or in the **entire** multiplex. Assuming 75-100 MB/sec throughput per core.
- **TEMP Store disk**: 2-3 drives per core on the host. Assuming 75-100 MB/sec throughput per core.
- **MAIN Store Fiber Controllers/HBAs**: 1 per 5-10 cores. Size based on total MB/sec throughput needed on host.
- **TEMP Store Fiber Controllers/HBAs**: 1 per 5-10 cores. Size based on total MB/sec throughput needed on host.

Other Resources

Contact me via email at mark.mumy@sap.com

Check out my blog: http://scn.sap.com/people/markmumy/blog

Use the IQ Community on the SAP Community Site:
http://scn.sap.com/community/sybase-iq

Use the IQ Users Group: iqug@iqug.org

IQ 16 Sizing Guide: http://scn.sap.com/docs/DOC-46166

Questions and Answers

Thank You for Attending

Please complete your session feedback form