SYBASE

SYBASE®
INFORMATION ANYWHERE.

Non-Stop Sybase IQ Technical Overview

Paul Krneta
Chief Technologist Sybase IQ
pkrneta@sybase.com

SYBASE

Agenda

- Non-Stop IQ Overview
- IQ Architecture Overview
- Non-Stop IQ Detail
- Examples

Why use NonStop Architecture ?

To solve the following problems:

- How do you backup a 10, 20 or 100TB of data while providing 24/7 access?
- How do you restore 10 or 100 TB DW without days-weeks of downtime ?
- How do you run maintenance on a 24/7 system?
- How do you guarantee 24/365 availability?
- 25%-40% of tape restores fail…is that acceptable ?
- What is the cost of multiple tape copies of 100 TB of data?

To benefit from low-cost disk ATA storage

- ATA cost per TB is the same as Automated Tape Libraries
- Raw HW-RAID ATA is < $2,000/TB, just like tapes
- ATA more robust, faster and usable that tape library

Answer: IQ Non-Stop Architecture

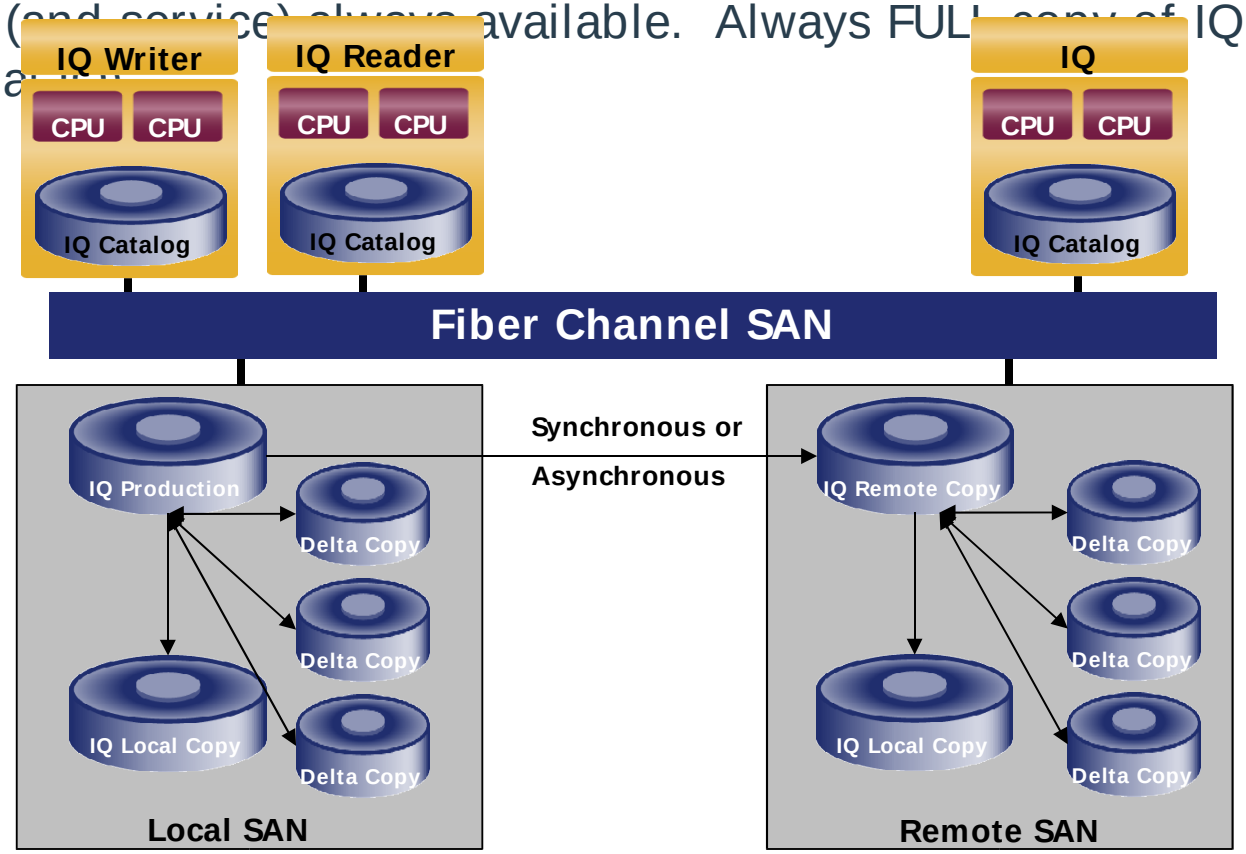
"NONSTOP IQ": SYBASE IQ WITH BUILT-IN BACKUP/ RESTORE, DR, 2nd SITE, FT, 24x7

- The Problem: Classic Backup technology can not handle Multi-TB DW:
  - Slow & Unreliable:
    - Backup & **Restores** take hours-days, impact production;
    - 24%-40% restores from tapes fail
  - Expensive Cost and complexity of 3rd party backup SW, costly operation, sunk cost
- **The Solution: NonStop IQ** : instant Disk-to-Disk (D2D) backup/ restore
  - **On-Line Backup & Zero-Downtime Restore**: **< 60 sec** using Disk-to-Disk (**D2D**) c
  - Local and **Remote** operations
  - **Active** Backup**:**
    - More reliable than tape backup: instant verification of a backup data
    - Backup can be used for Development, testing, QA etc.
  - **Active Backup enables active DR site**
  - Lower cost than "classic" tape Backup & Disaster Recovery
- 10TB DW Certification Sybase IQ+ Sun+ EMC compl
- +20 customer have it in production: DW size from 200GB to 40 TB

Non- Stop IQ

"Non- Stop IQ" is a term that describes how to integrate Sybase IQ VB/ VR technology with data services (provided by storage array or host).

This combination can be used to create HA/ DR configurations that can make your data (and service) always available. Always FULL copy of IQ (never incremental!!!)



Non- Stop IQ – Full Copy of MainIQ

Synchronous Copy or Mirrored Copy

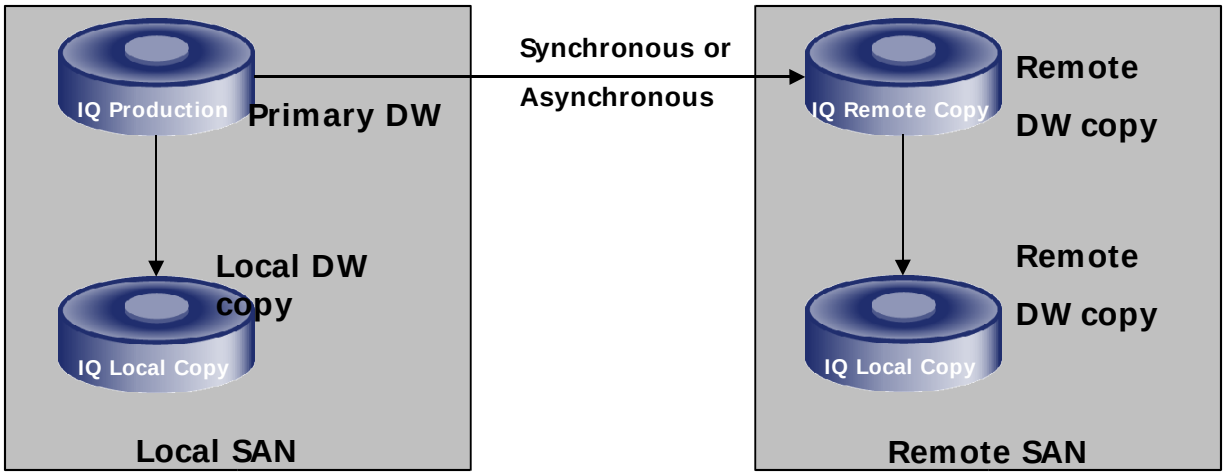
One or More copies of a single source
A mirrored copy can be broken or synchronization stopped instantly
- Normally a bit map is used to track changes
- Changes can be resynchronized quickly

Remote Synchronous Copy

Usually requires high speed link with dedicated bandwidth
One of the best solution for continuous availability configuration

Copy means a Full copy versus Delta copy

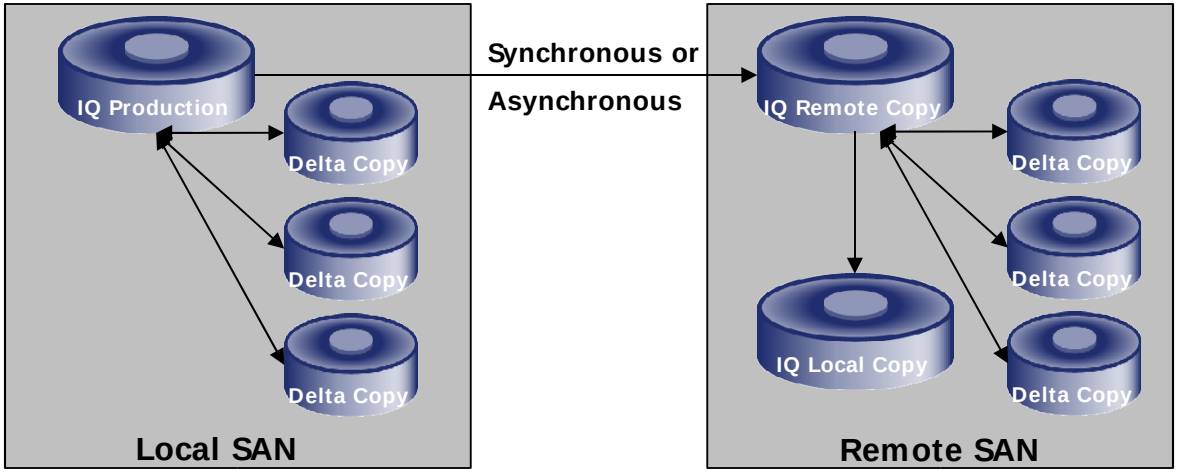
Requires equal amount of disk space for the copy
Can take significant amount of time to create initial copy



Non- Stop IQ – SAN Delta Copy

Delta copy creates a point in time copy of the database. This is otherwise known as a "copy on first write" or SnapShot copy.

- Every sector that is changed after the snap is taken, is copied, before the first time it is changed, to a separate device. Think of this as "redo log" .
- SANs allow the OS to mount the snapshot as if it were a full copy.
- Uses very little disk space. (e.g. If you have 5 years worth of data and you do a full copy everyday and a snap every hour. An hours worth of changes is very little when compared to 5 years of data.)
- Allows customers to start separate instances of IQ on copies without corrupting full copy. Can be used for maintenance, development or any other functions.
- Some SANs can rollback the full copy using the delta copy.
- SANs allow as many as eight delta copies from one source.



Non- Stop IQ SAN Scripts

For a particular SAN the customer will need to create scripts for:

Local Copy

create_localcopy – creates a full copy of the IQ Main devices on a the local SAN.
suspend_localcopy – suspends the mirror synchronization while keeping change list
reverse_sync_localcopy – reverse synchronizes local copy (copies “backup” to primary)
drop_localcopy – removes local copy.

Delta Copy

create_deltacopy – creates a delta copy or snapshot of the IQ devices on a the local SAN.
rollback_deltacopy – rollback source devices to this point- in- time image (w/ or w/ o changes to it)
drop_deltacopy – removes delta copy.

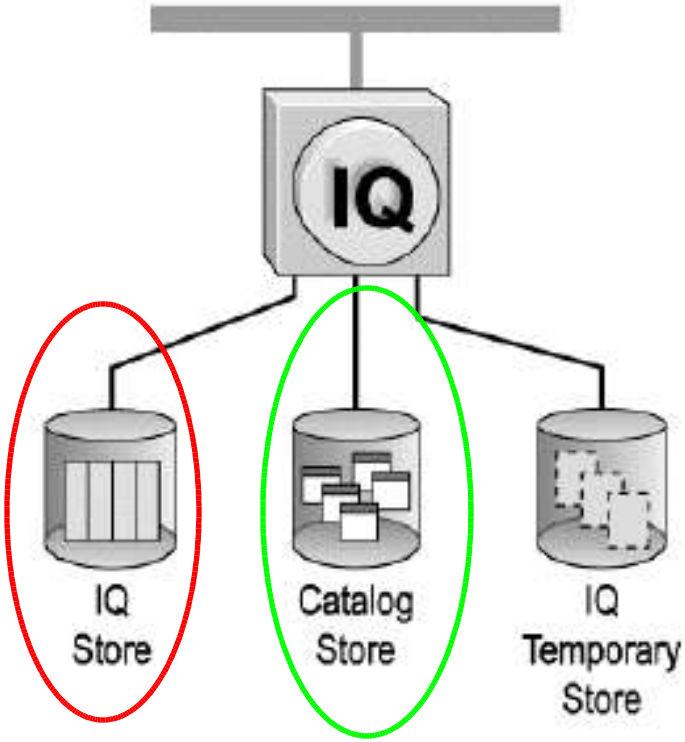
Remote Copy

create_remotecopy – creates a full copy of the IQ devices on a the remote SAN.
stop_remotecopy – breaks the mirror/ stops synchronization.
sync_remotecopy – synchronizes remote copy
reverse_sync_remotecopy – copies remote “backup” copy to primary copy
drop_remotecopy – removes remote copy

OR use host- based DD command (practical for up to 1TB)

3 main components of IQ Architecture

1 node (Simplex)



IQ Store ("MainIQ") and Catalog Store define IQ instance

(TempStore can be dropped/ added without any consequences)

Non- Stop IQ: Virtual Backup

Virtual Backup Command

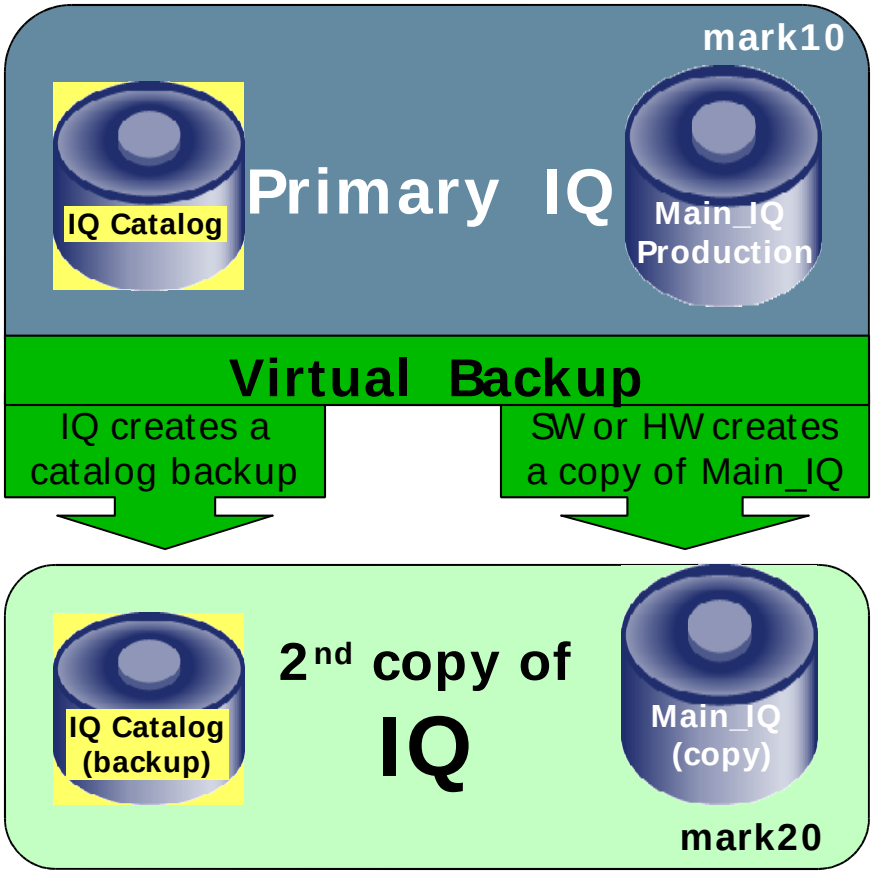
b. Creates a copy of the IQ catalog (as of start time of VB)

c. Runs script that uses storage commands to create a copy of the IQ data OR break a mirror at a storage consistent time.

d. No downtime

e. Backup copy can be used

This IQ Catalog and Full copy of MainIQ are consistent copy of IQ at the time when VB was started. VB took seconds to do it and complete and caused no downtime.



IQ Virtual Backup (VB)

VB: coordinates no-downtime IQ "micro-quiesce" with SAN command:

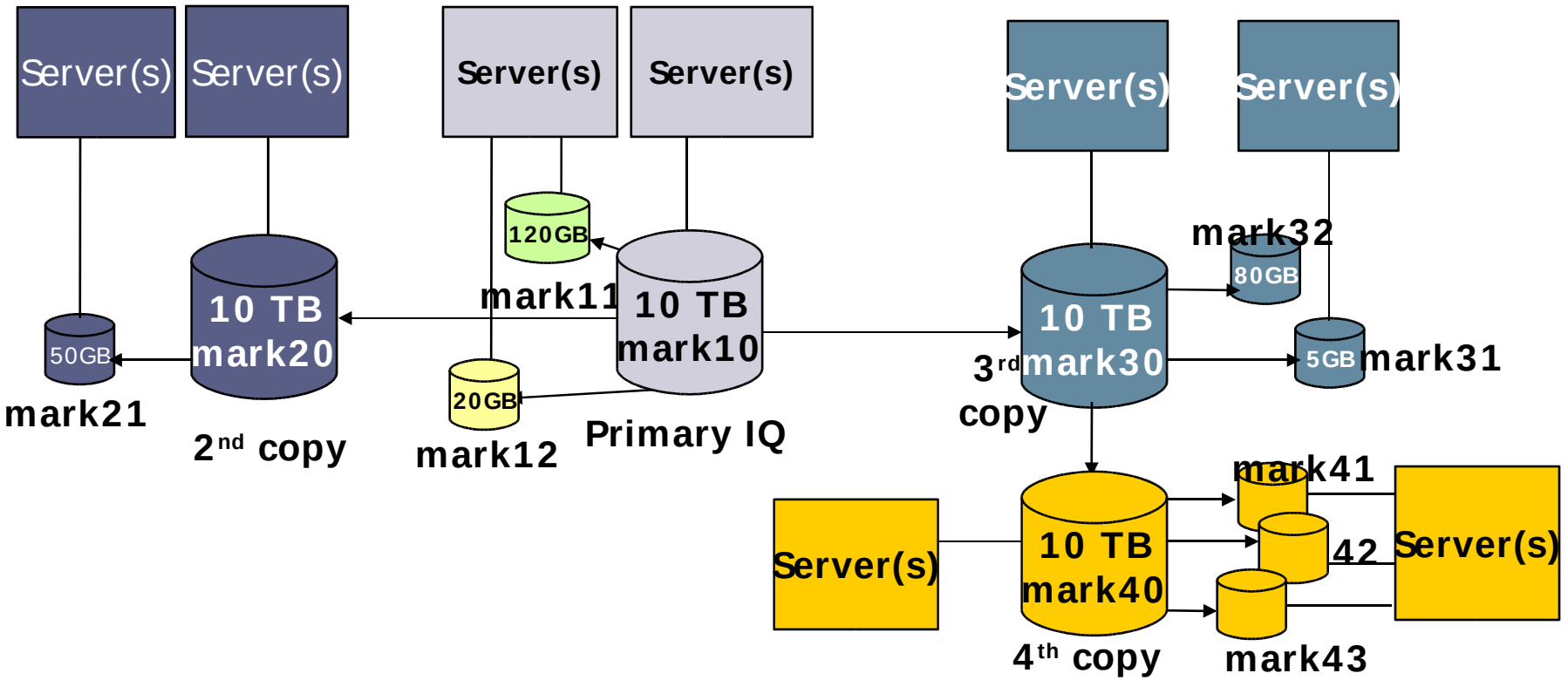
The IQ VB command does the following :

1. first creates a backup of IQ catalog,
2. calls script that creates a copy of MainIQ by using and coordinating SAN technology (or "dd") to copy, synchronize, or break a mirrored copy of a consistent copy of the database. When VB finishes the copy of IQ catalog and MainIQ are there. The VB takes < 60sec without stopping IQ.

SAN replication features allow synchronous and asynchronous, local and remote, as wells as full copies or delta copies of the database devices.

The combination of these features can be many and they have met all customers HA, DR or CA needs….

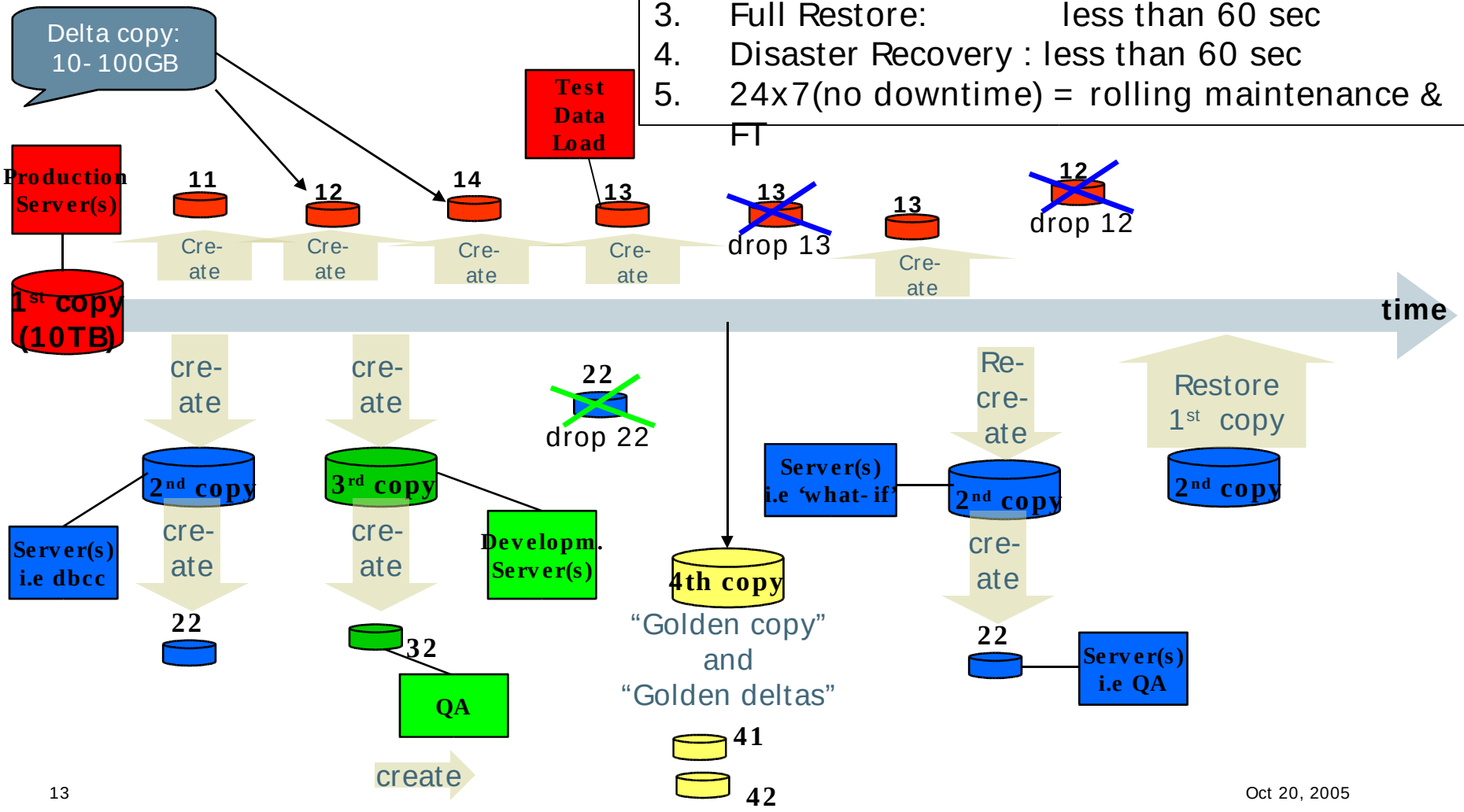
"NONSTOP IQ": SYBASE IQ WITH BUILT- IN BACKUP/ RESTORE, DR, 2nd SITE, FT, 24x7



NonStop IQ : Operational possibilities

All Backup/ Restore operations are instant:

2. Create backup : less than 60 sec
3. Full Restore: less than 60 sec
4. Disaster Recovery : less than 60 sec
5. 24x7(no downtime) = rolling maintenance & FT



SYBASE

Non- Stop IQ
Other Non- Stop Tips

Create a Naming convention and use a simple number mechanism to identify full copy or delta copy LUNs for each instance:

Use Instance name plus 2- digit- number. First digit defines full copy and second digit denotes delta copy from that full copy:

i.e. system name = Mark

Mark10 is the primary production copy

Mark11 is the first delta copy of Mark10

Mark12 is the second copy of Mark10

Mark20 is a full copy of Mark10

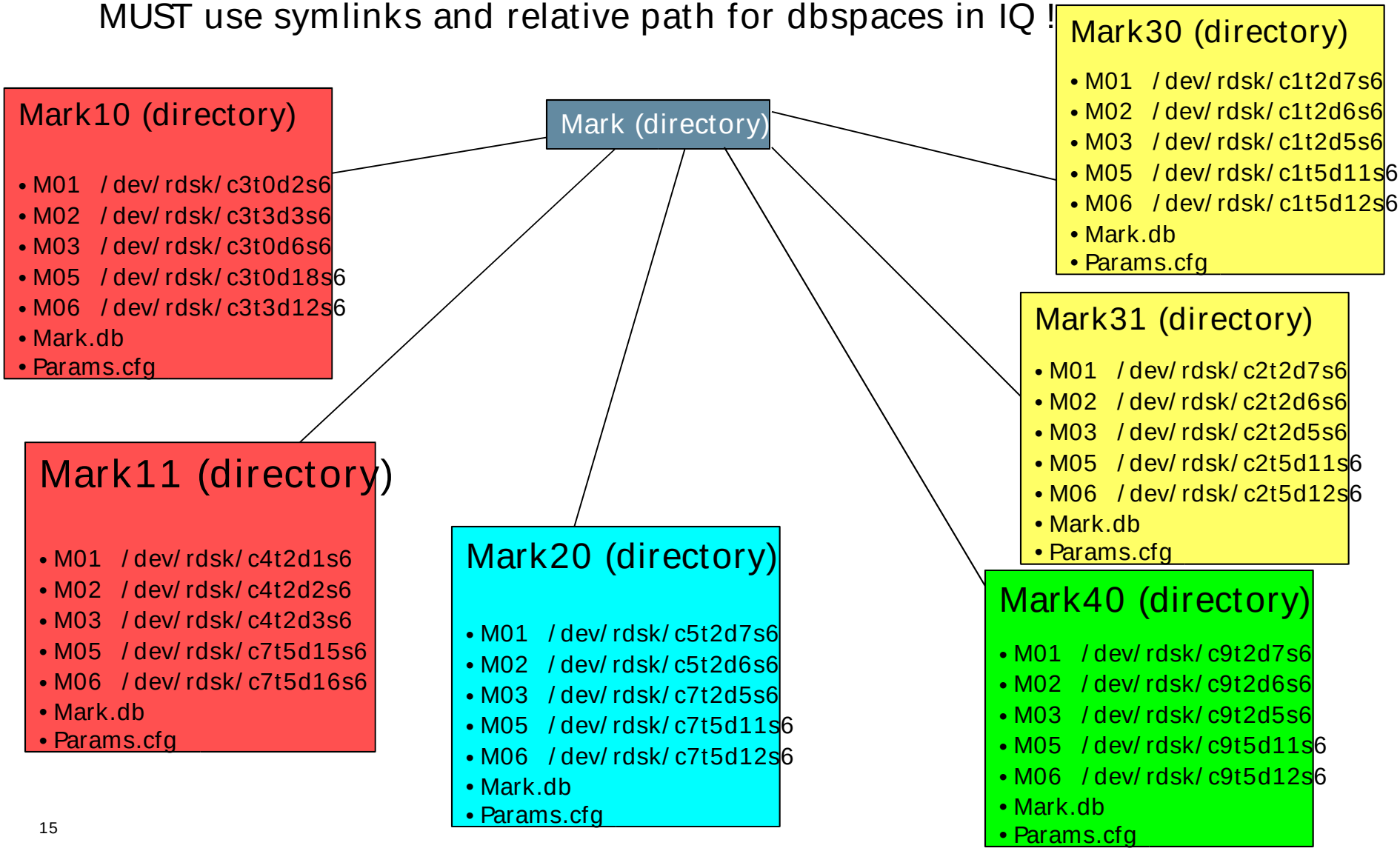
Mark21 is the first delta copy of Mark20

Mark30 is the third full copy of Mark10

Etc…

NonStopIQ Directory Structure

MUST use symlinks and relative path for dbspaces in IQ !



How can you afford multiple copies of 100TB of data?

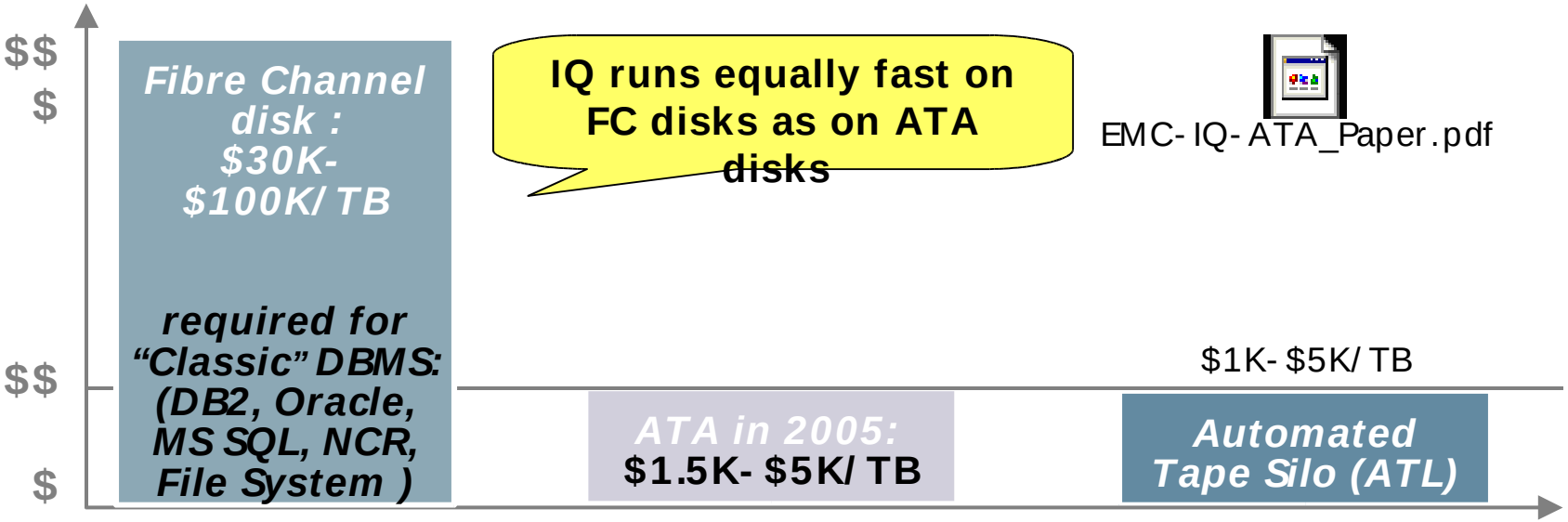
Reduce the size of the database so you can fit 3 or 4 copies using the same storage size [in TB] (or 2 copies using half the storage size)

- Standard DBMS explode data via indexes and aggregate tables
  - 30TB of raw data will routinely explode to 90TB with indexes
- IQ stores the data by column, everything is indexed and compressed
  - IQ loads 30TB of data into a fully indexed 15-25TB database

Leverage less expensive Serial ATA disks…

- IQ accesses disk very differently than other databases. Large IOs only…
  - Other databases are designed to process records therefore they do many small reads. IQ is designed only for analytics and therefore does only large reads of each columns bit wise index….
- IQ therefore gets tremendous performance out of less expensive Serial ATA disks
- IQ I/O pattern looks like “video streaming”: large[150KB-512KB], semi-sequential
  - Use larger IQ page size (256KB or 512KB)

IQ+ ATA : +90% LOWER STORAGE COST THAN "Classic" DBMS – SAME COST AS TAPE SILO



IQ: ¼ of size of "Classic"

X

¼ of price per TB (ATA)

= 90% storage savings

Performance sizing:

- **Min. 1 ATA DISK per IQ CPU_core**
- **Min. 0.5 FC DISK per IQ CPU_core**

Example of ATA SAN Pricing (Oct 10 2005)

www.sun.com :

StorEdge 3511 (HW RAID, dual-controller, +350 MB/sec)

Using 400 GB ATA disks

28.8 TB SAN for $98,000 $3,400 per TB

www.apple.com

Xserve RAID :

7TB for $13K $1,857 per TB

IQ: USING HIGH-DENSITY STORAGE

4 TB of usable (formatted) capacity [IQ will fit 5-10 TB of data]:

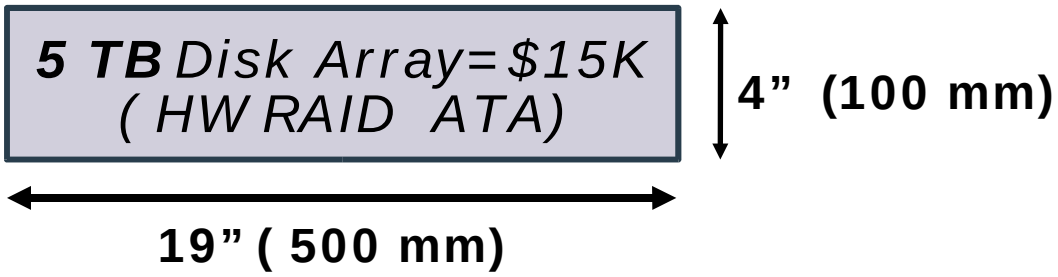
- 100 Billion records (100 bytes each)
- 5 Million documents (1 MB each)
- 1,000-2,000 movies (3 GB each)
- 100,000 songs (50MB each)

IQ runs on ATA disks at the same speed as on FC or SCSI disks

Price= $15K $2K-$5K/TB same or lower cost than tape silo or optical

Read: 400 MB/sec [1.3 TB/h] 1,500 Hi-Def video streams or 400 docs/sec

ATA Storage vendors: EMC (AX100), Sun (3511, 6130), HP, IBM…



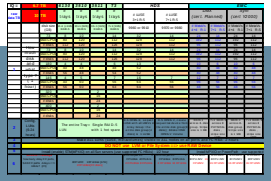
SYBASE®

How many CPUs for IQ ?
How much storage for IQ ?

	IQ	"Classic" DBMS (Oracle, DB2, MSSQL, NCR)	IQ advantage
How many CPUs ?	1) 25% fewer than existing (non-IQ) DW or	more than IQ	big
What type of servers	Flexible, not critical	large SMP	big
RAID type	RAID-5 (or RAID-3)	RAID-1	80%
Parallel DW (multi-node) ?	Yes (Multiplex): start with "Mpx-in-the-box"	complex	big
DW_size(logical) relative to input data	0.3x-0.9x of input data	3x-10x input data	3x-10x
Raw Storage size	IQ_DW_size + 15%_overhead_for_R-5	Dbsize+100%(R1)	
DB block size (unit of I/O)	512KB-256KB	4KB-32KB	8x-128x
Required IO bandwidth per CPU core	10MB/sec	>100MB/sec	6x
I/O BW storage server required	#cores X 10 [MB/s]	#CPUx100 [MB/s]	6x-100x
Required IOPS per CPU	50	>1,000	20x
Minimal # of spindles per CPU	0.5 (FC) to 1 (ATA) per CPU_core	6-10	6x-30x
Max_Num_CPUs per FC (2 Gbps)	8-16	1	8x-16x
Optimal disk size (GB)	146GB, 250GB-320GB-400GB ATA..)	Small (36-146GB, FC)	2x-10x
Enough IO bandwidth from load files ?	Provide high BW from load files to IQ		

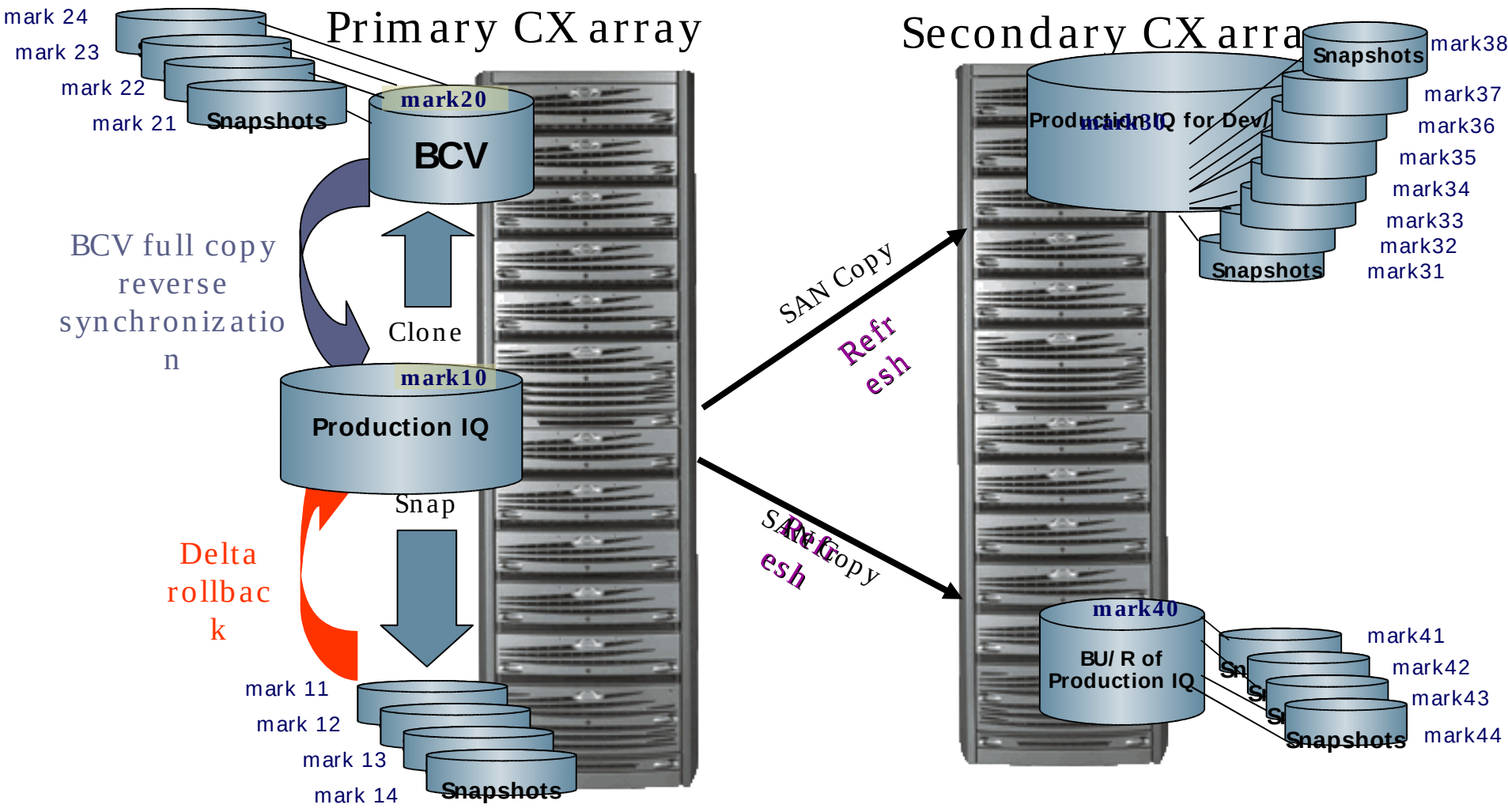
Building Optimized Storage Configuration for IQ : 9 steps

SYBASE

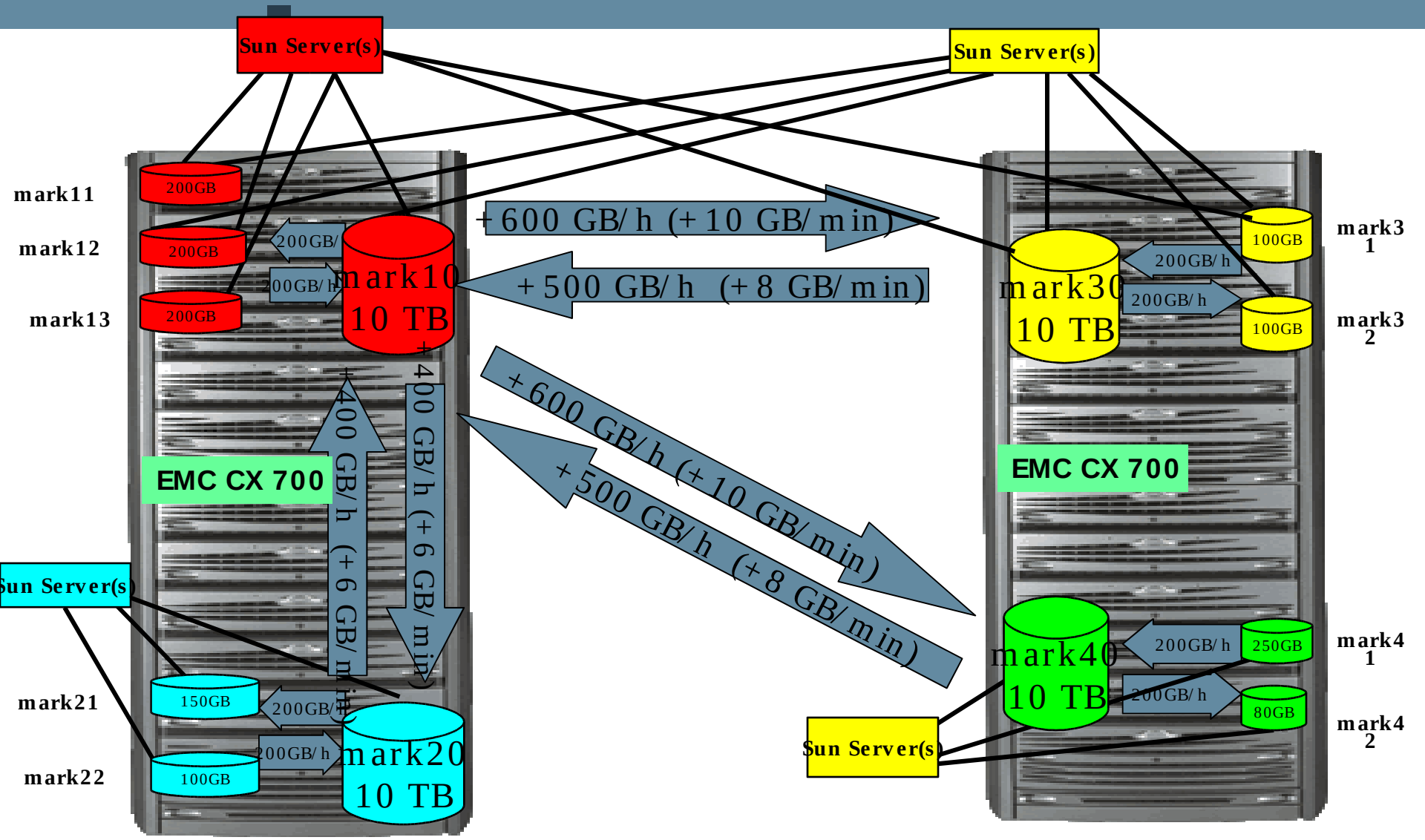


IQ =	6.7 TB		6130	3510	3511	T3	HDS		EMC					
raw data TB	10 TB		# trays	# trays	# trays	# trays	# LUSE 3+1 R-5	# LUSE 7+1 R-5	DMX (cert. Planned)		Sym (cert Y2000)		Clariion CX: see slides	
		disk size (GB)	12+1 (+1HS) RAID-5	10+1 (1HS) RAID-5	10+1 (1HS) RAID-5	7+1 (1HS) RAID-5	9960 or 9910	9970 or 9980	# MetaVs 4+4 R-1	# MetaVs 7+1 R-5	# MetaVs 4+4 R-1	# MetaVs. 7+1 R-S	ATA 4+1 R-3	FC 14+1 R-5
		73	8	10		14	32	14	24	14	24	14		
		Max CPUs	80	120		112	128	112	192	112	192	112		
		# disks	112	120		126	128	112	96	112	96	112		
	How much disk and what type ? (1 hour)	73	8	10		14	32	14	24	14	24	14		
		Max CPUs	80	120		112	128	112	192	112	192	112		
		# disks	112	120		126	128	112	96	112	96	112		
		180	4	4		6	13		10	6	10	6		
1		Max CPUs	40	48		48	52		80	48	80	48		
		# disks	56	48		54	52		40	48	40	48		
		146	4	5		7	16	7	12	7	12	7		4
		Max CPUs	40	60		56	64	56	96	56	96	56		56
		# disks	56	60		63	64	56	48	56	48	56		60
		320			2								6	
		Max CPUs			16								24	
		# disks			24								30	
		400			2								5	
		Max CPUs			16								20	
		# disks			24								25	
2	Config. LUNs (6-24 hours)		The entire Tray= Single RAID-5 LUN with 1 hot spare				R-5 OPEN-E, sequen. LUSE with all LDEVs in Array Group: the entire disk group(4 disks) is 1 LUSE	R-5 OPEN-V, create sequential device from the entire disk group(8 disks); 8disks= ONE OPEN-V Volume	RAID-1 across 8-disk group. Stripe size is 1 MB	7+1 RAID-S across 8 PHYSICAL disks , stripe size	RAID-1 across 8-disk group. Stripe size is 1 MB	7+1 RAID-S across 8 PHYSICAL disks , stripe size		
3	Make ALL LUNs (LUSE, metavolumes) visible to ALL nodes on all ports and FC HBA: 2 hours													
4	DO NOT use LVM or File System ==> use RAW Device													
5	Install (enable) STM(MPXIO) on all Sun servers (use supported FC HBAs) : 1/2 hour								Install MPXIO or PowerPath : use supported FC HBA : 1hour					
6	Servers<=>Storage how many Array FC ports, SAN FC paths Arrays <=> HBAs? (2 h)		8CPU/FC 1CPU/disk [ATA] 2CPU/disk[FC disks]				8 CPUs per FC 1 CPU/disk 4CPU/AG 2 CPU/disk 8 CPU/AG	8 CPUs per FC 1 CPU/disk 8CPU/AG 2 CPU/disk 16 CPU/AG	8 CPU/FC 8CPU /MV 16 CPU/MV.	8 CPU/FC 8CPU /MV 16 CPU/MV.	8 CPU/FC 8CPU/MV 16 CPU/MV	8 CPU/FC 8CPU/MV 16 CPU/MV.		

Sybase IQ / EMC CLARiiON: Example #1



IQ- EMC Validation : Performance of background data copy



SYBASE

Non- Stop IQ
Virtual Backup Example continued..

Virtual Backup Procedure other points:

How often do you load data?

IQ backup after every IQ load, or

Copy IQ load files to DR site to be rolled forward in case of Disaster. After re-synchronizing remote copy, delete unnecessary load files.

While re- synchronizing the local copy (on previous slide) and before the Virtual Backup command is run, the local copy is no longer valid. Solutions:

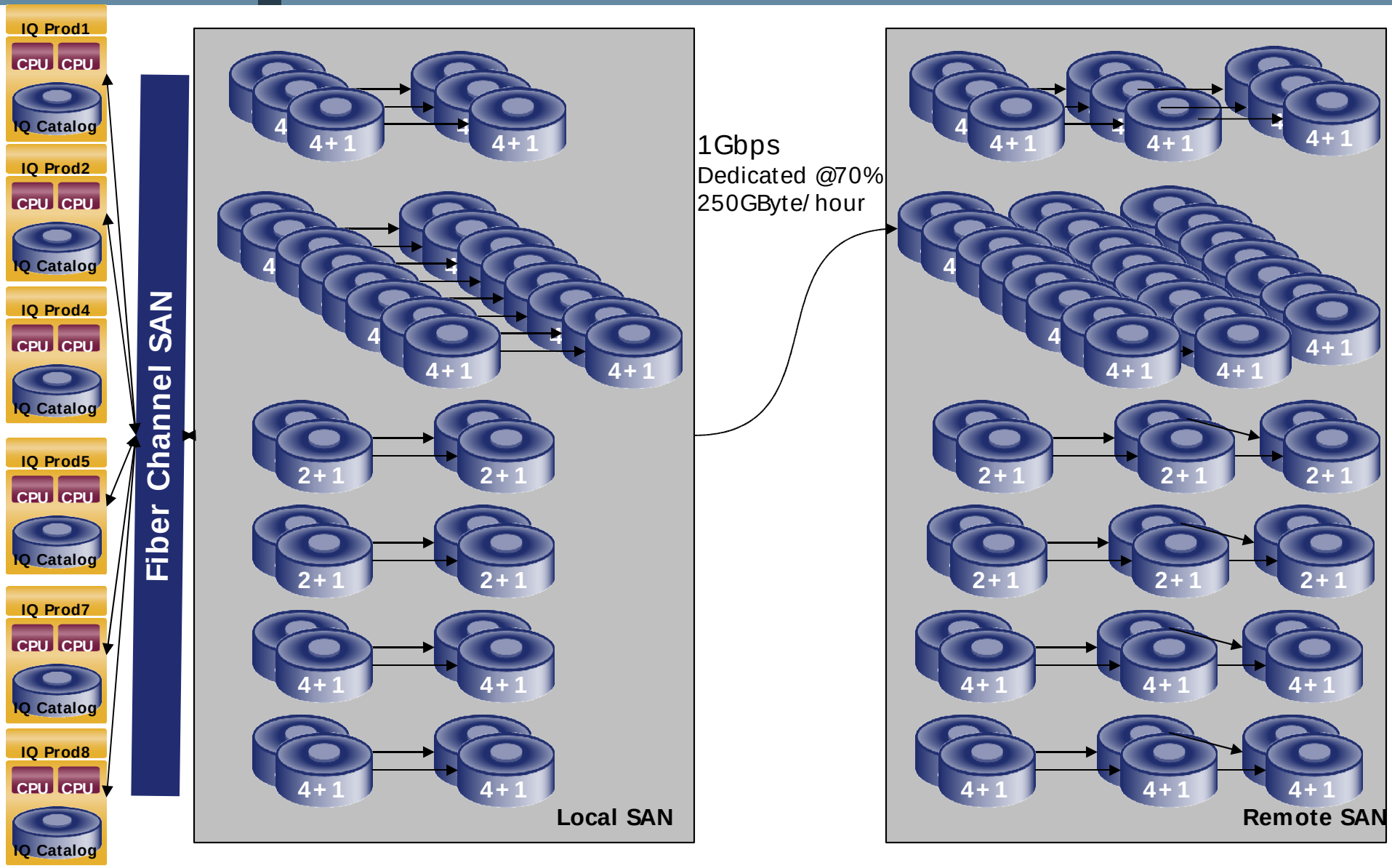
Have a third or gold copy (Remote copy in previous slide), or

Create a snapshot copy or delta copy that can roll back the local copy, before starting the re- synchronization.

99% of all IQ databases use IQload to load data but if you are using inserts, updates and deletes...

Create a local or remote synchronous copy and use IQ multiplex. Multiplex instance can be on the same or different system. Different system can be local or remote. Two or more copies of the data are always better one.

Customer Example



Customer Example
Total Disks

Chicago CX700		Primary copy			2nd copy		RFC CX700						
System	CPUs	Total Disk (TB)	LUNs	Disks	Mirrored	Tota Size (TB)	System	CPUs	Total Disk	LUNs	Disks	QA	Dev
Prod1	10	3.84	3	15			Prod1		3.84	3	15	15	15
Prod2	4	10.24	8	40			Prod2		10.24	8	40	40	40
Prod4	6	1.28	2	6			Prod4		1.28	2	6	6	6
Prod5	8	1.28	2	6			Prod5		1.28	2	6	6	6
Prod7	6	2.56	2	10			Prod7		2.56	2	10	10	10
Prod8	4	2.56	2	10			Prod8		2.56	2	10	10	10
Total	38	21.76	19	87	174	**43.52**	Total				87	87	87

174 SATA Disks (primary and secondary copies)
14 FC Disks for reserve LUNs (delta copies)
188 Total Disks of 240 Possible (73%Used)

***Would require use of existing cx600s

SYBASE®

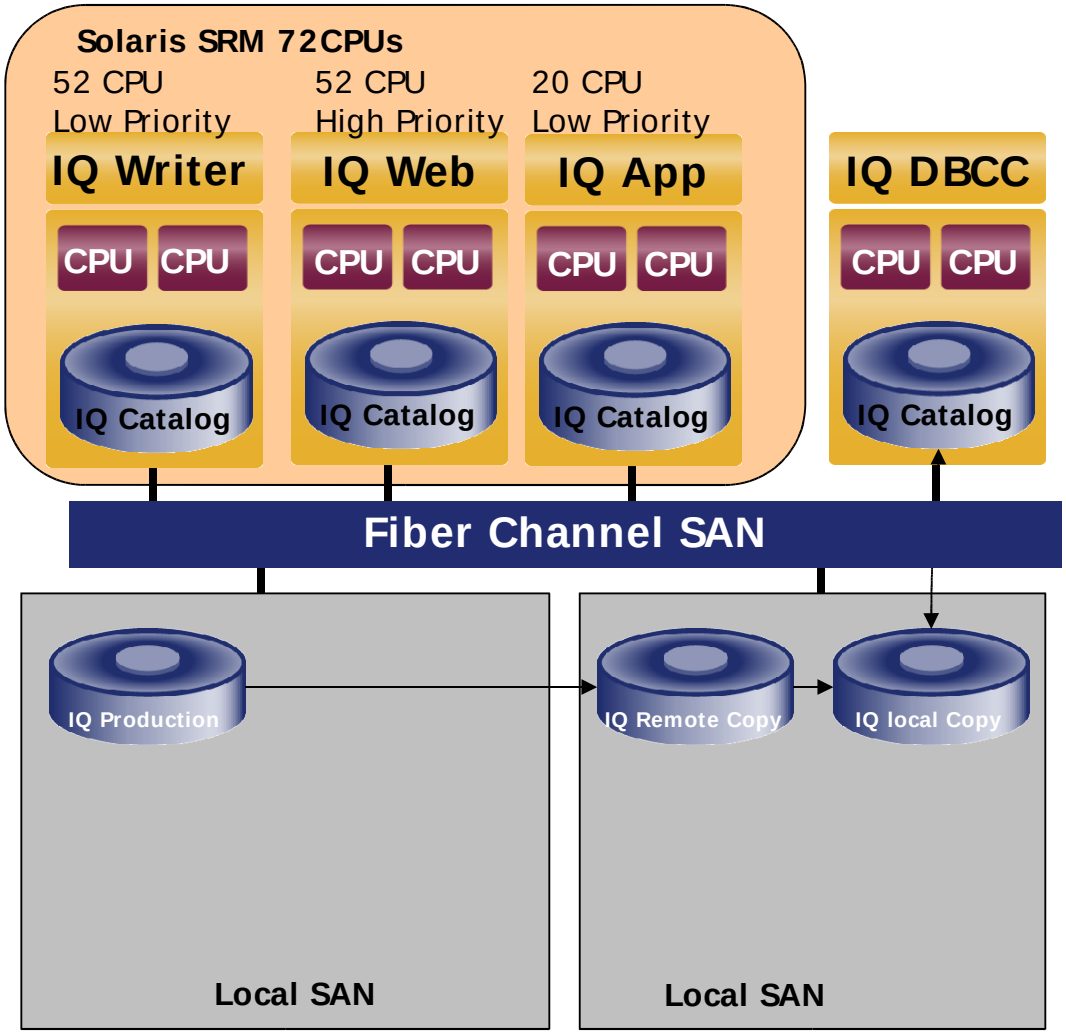
Questions ?

SYBASE®

Thank you

How do you run maintenance on a 24/7 system?

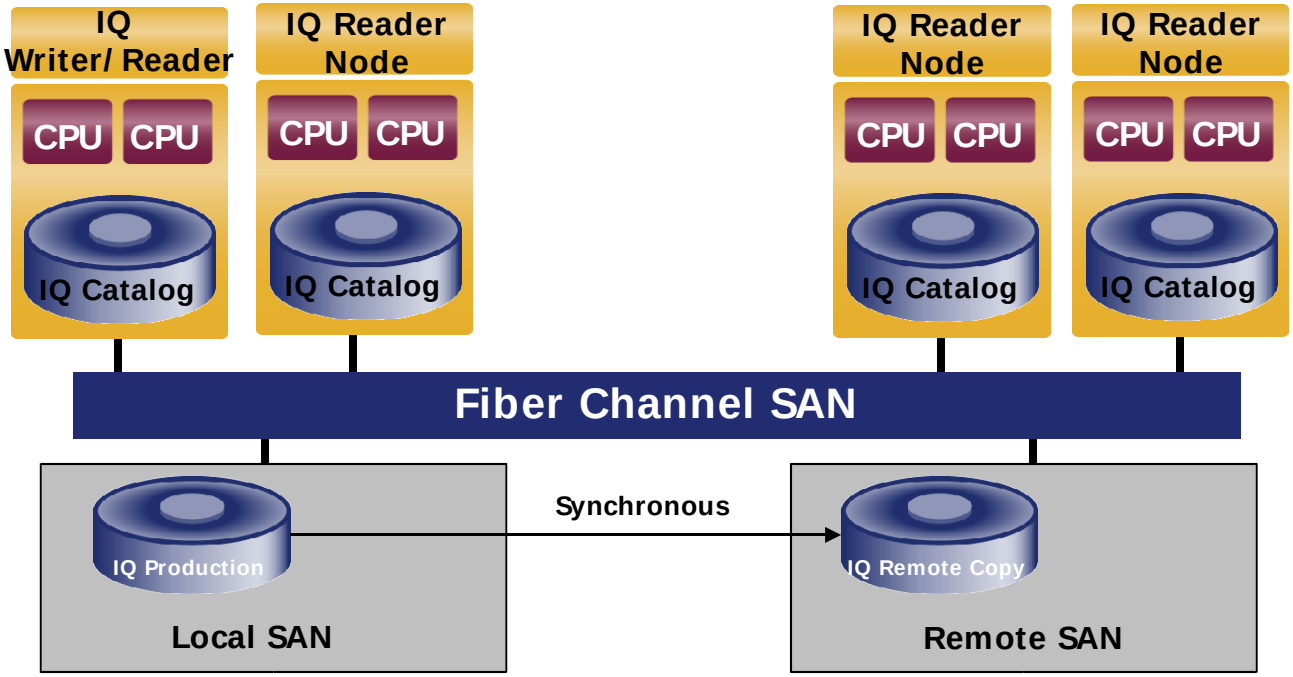
Run consistency checks and other checks on a separate system using a copy of the database...



availability?

By having multiple copies of the database...

Access the copies simultaneously or start remote on local failure, depending on your maximum failover time...



database while providing 24/7 access?

Leverage SAN technology that can create copies immediately

- SAN technology can create multiple mirrors or copies of the data…
- These copies can be local or remote…
- Sybase IQ virtual backup command is integrated with the SAN mirroring scripts…
- Sybase IQ virtual backup command can create backups instantly no matter how large the database.

Reduce the size of the database so that you can fit 3 or 4 copies using the same number of disks or 2 copies using half the disks…

- Standard databases explode data creating indexes and aggregate tables
- A standard database takes 30TB of raw data and explodes it to 90TB with indexes
- Sybase IQ stores the data by column in bit wise indexes, therefore everything is indexed and compressed.
- Sybase IQ loads 30TB of data into a fully indexed 20-30TB database

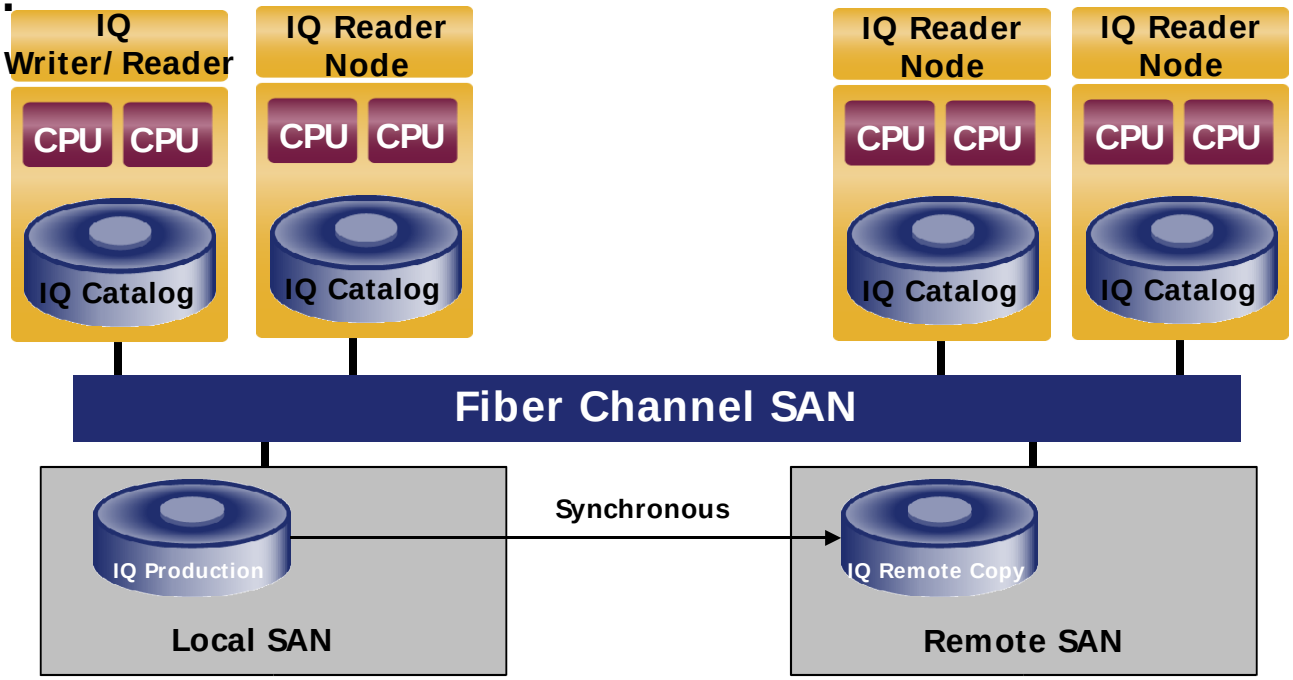
Non- Stop IQ Synchronous Remote Copy

A SAN remote synchronous copy can create a copy at the remote site and keep it in sync synchronously.

Local instances of IQ Multiplex on local systems use local database devices and remote instances of IQ Multiplex on remote systems using remote database devices.

IQ Multiplex will keep IQ catalogs in sync across all of the systems.

A remote reader node can be promoted to a writer node if the local site fails.



SYBASE

Non- Stop IQ
Other Non- Stop Tips

Keep CPU to Disk Ratio of around 1 to .5- 1 (total number of CPUs across all nodes of multiplexed configuration)

- IQ does many large IOs (128K default page size) therefore it runs just as fast on less expensive large SATA disks ($1.5K- $5K/ TB) RAID5. (RAID3 on Clariion)
- Use at least a dual fiber channel SAN with Highly Available host software that can re-route around a failed host controller.
- A typical 2 fiber channel SAN can maintain 200- 400MB/ second throughput.

IQ likes a few large devices – easy administration

- Most SANs have a 2TB logical RAID device limit.
- Use entire logical RAID device e.g. 4+1 RAID3/ 5 with 320GB approx 1.2TB.
- 320GB SATA disks will take a while to rebuild after a failure, so therefore don't go to 8+1 or larger RAIDs to lower the probability of a second disk failure while syncing.

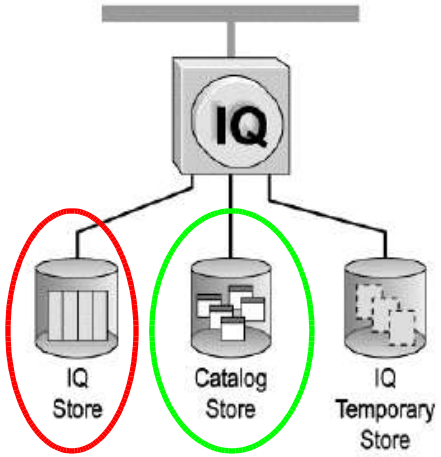
Limit 8- 10 CPUs per Fiber Channel

Create dbspaces with expandable option

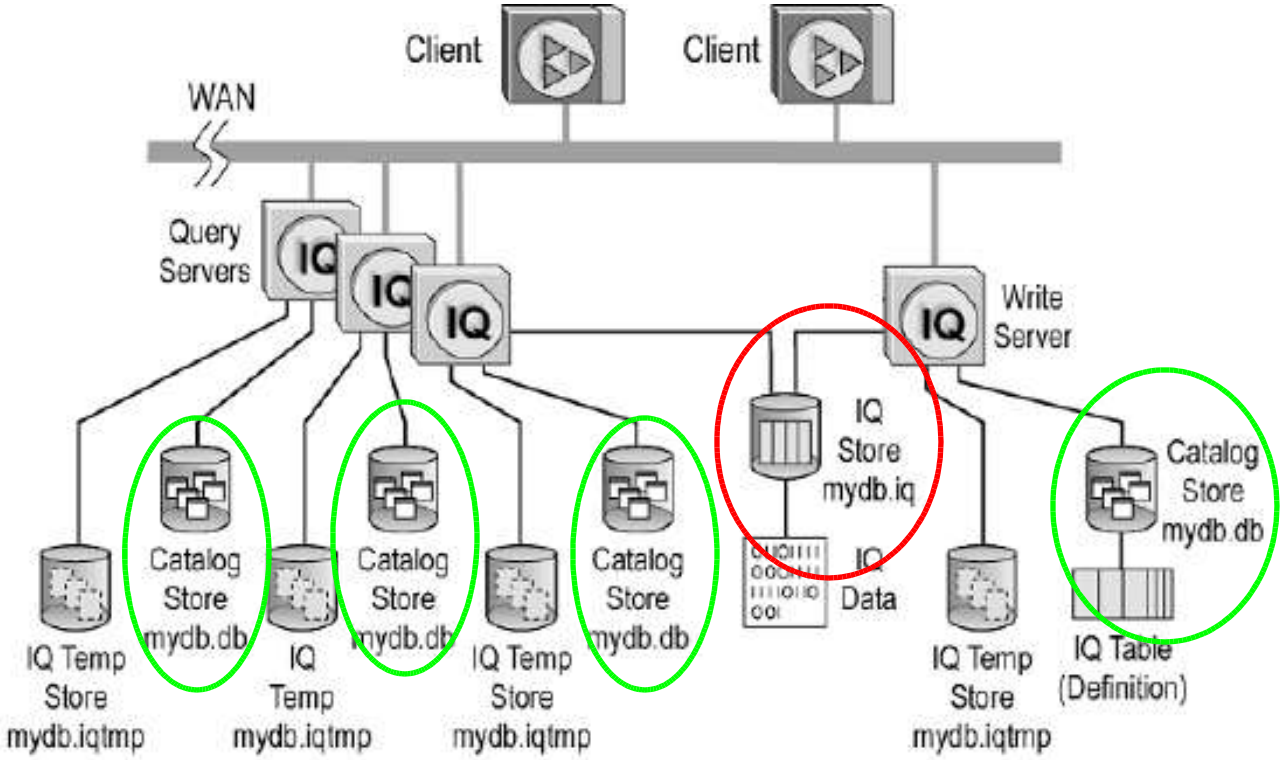
MUST Symbolic links Relative path for dbspaces

3 main components of IQ Architecture

1 node (Simplex)



IQ Multiplex

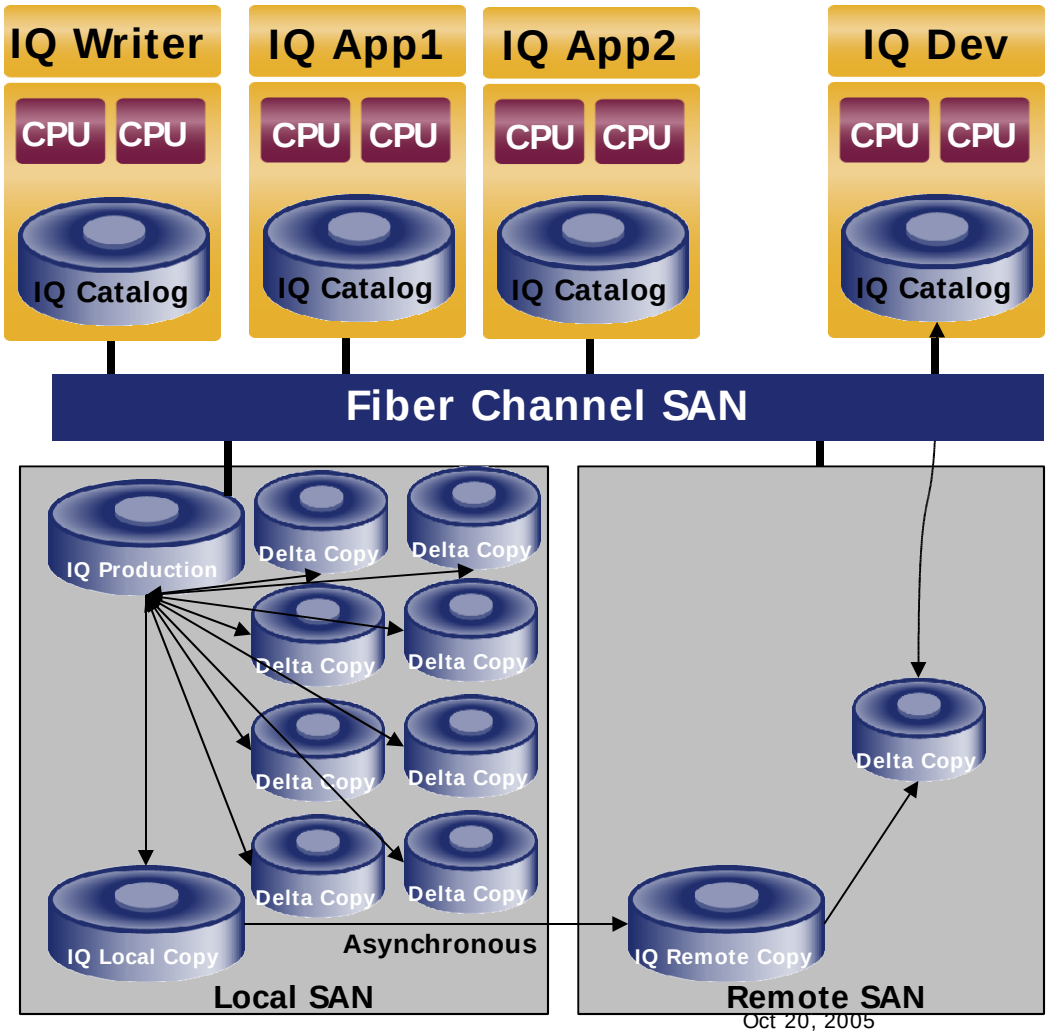


Each server has its own IQ Temp and Catalog Store and shares the IQ Store

Customer Example 1

Example 1:

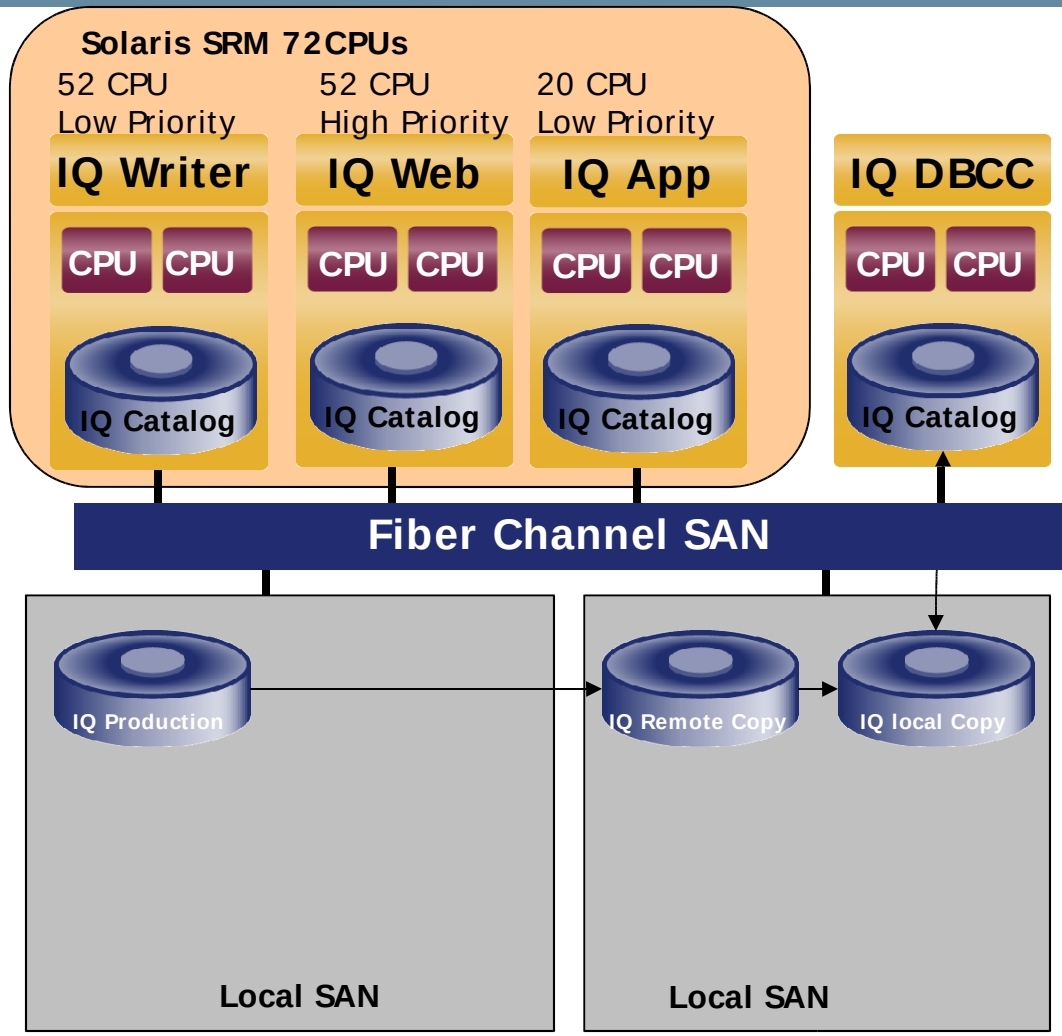
- 3 node multiplex
  - Loads won't effect queries.
  - App1 queries won't effect App2 and visa versa.
- Daily re- synchronization of full copies local and remote.
- Hourly delta copies during business in case of operational errors.
- Remote delta copy for development will not effect DR remote copy.



Customer Example 2

Example 2:

- 3 node multiplex
  - Loads are low priority therefore only use idle processors.
  - App queries won't effect Web queries and visa versa.
- Daily re- synchronization of full copies.
- Remote second copy for database consistency checks.



Files Allocated When Creating Database

Default dbspaces created with Create Database

dbname.iq* – IQ Main Store contains IQ Data and Transaction Log (raw device or file system)

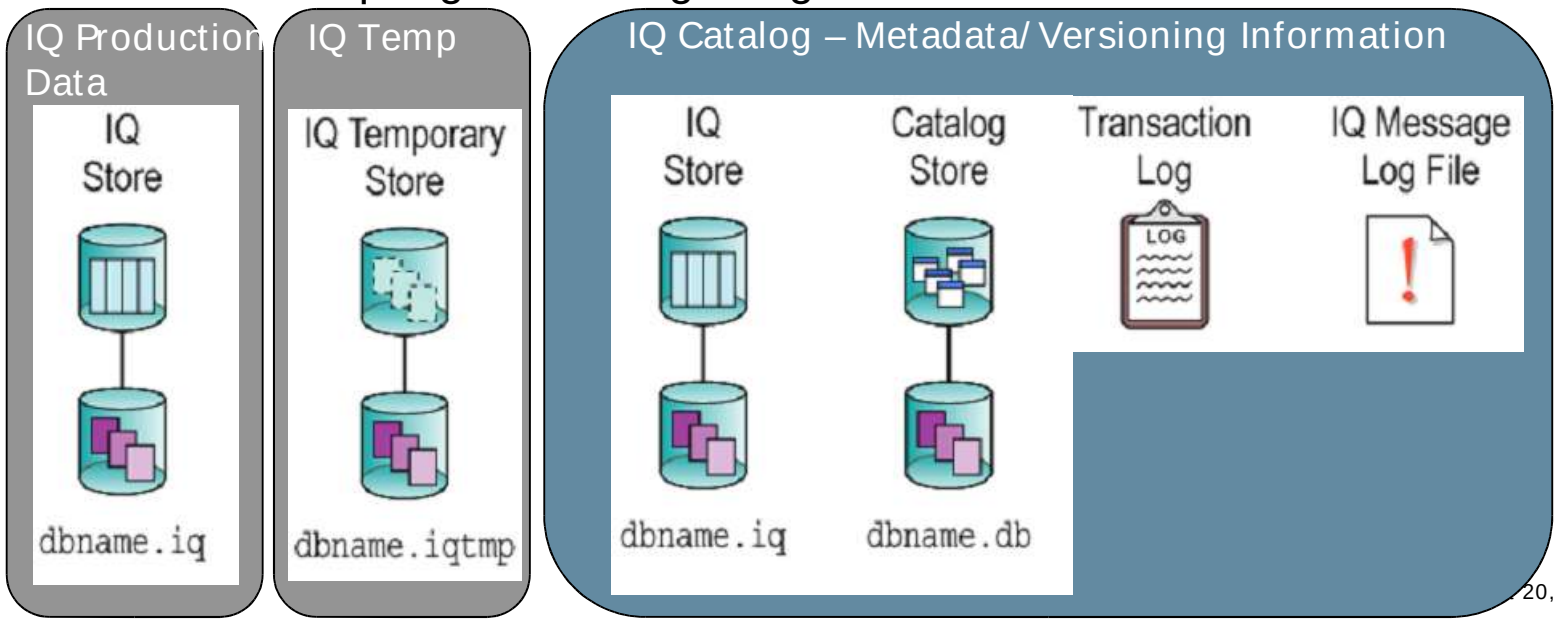
dbname.db – IQ Catalog Store contains IQ metadata

dbname.iqtmp* – IQ Temporary Store (raw device or file system)

Other objects created with Create Database

dbname.log – IQ Catalog Transaction Log

dbname.iqmsg – Message Log



Non- Stop IQ Virtual Backup Example

Virtual Backup Procedure

First time only:

- Create initial copies.
- Run Virtual Backup command to break mirror and create a catalog backup.
- Resynchronize remote copy.

Other times:

- Resynchronize local copy.
- Run Virtual Backup command to break mirror and create catalog backup.
- Resynchronize remote copy.

Virtual Restore Procedure

- Stop instances of IQ running on Production copy.
- Restore catalog from backup.
- Point IQ devices to local copy (now production copy) and start.
- Recreate local copy.