Namik Hrle
IBM Distinguished Engineer
hrle@de.ibm.com

Scalable Analytics: IBM System z Approach

Symposium on Scalable Analytics - Industry meets Academia
FGDB 2012 FG Datenbanksysteme der Gesellschaft für Informatik e.V. (GI)

November 22-23, 2012
Garching bei München

Disclaimer

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Enterprise Information Is on System z

66 out of the top 66 banks in the world

9 of the top 10 global life/health insurance providers

24 of the top 25 US retailers

IBM DB2 – Up to 4.8M connected threads

TD Bank – Zero unplanned downtime for over 10 years

Large U.S. Retailer – World's largest SAP system

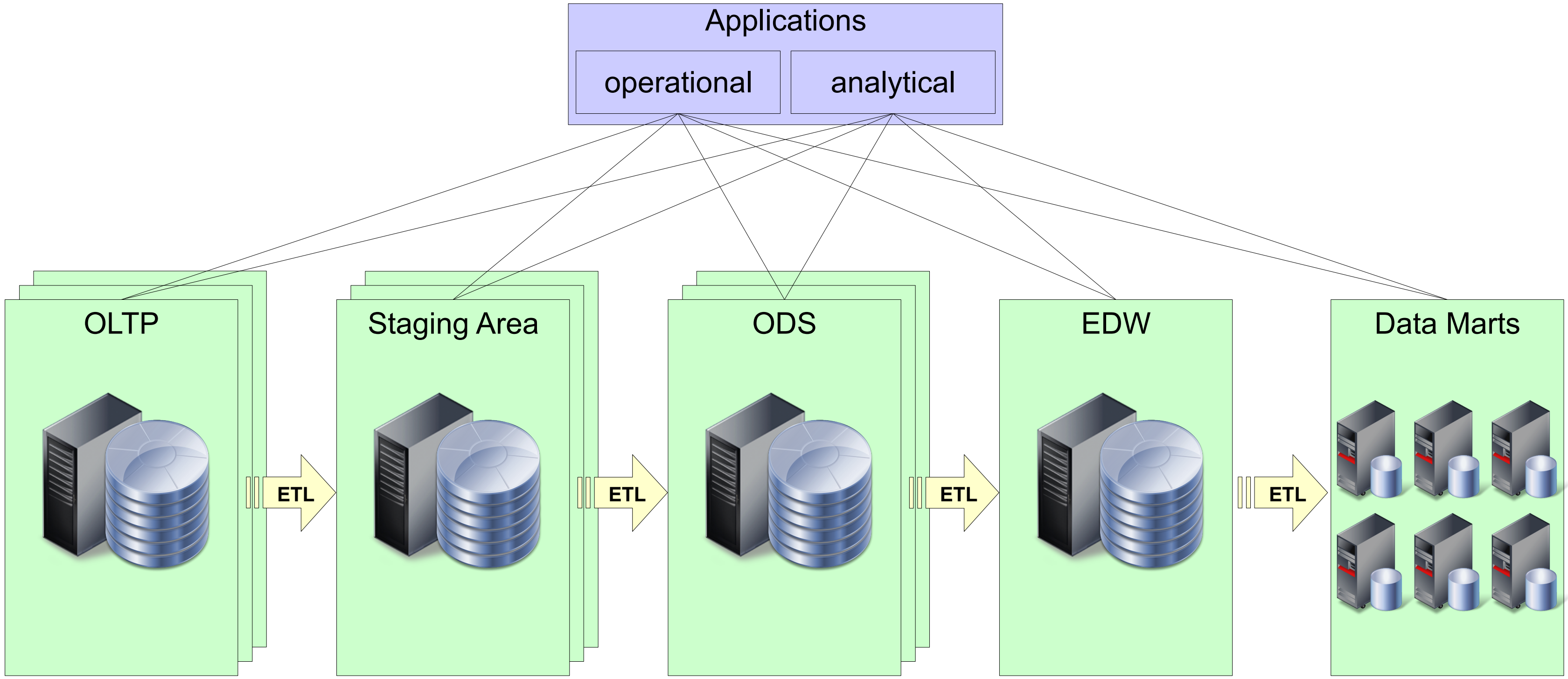
UPS – World's largest known peak RDBMS workload 1.1 Billion SQL statements per hour

ICBC 中国工商银行 INDUSTRIAL AND COMMERCIAL BANK OF CHINA – 620M accounts, 120M online users

中国银行 BANK OF CHINA – World's largest banking benchmark result, a record 9,445 transactions per second

Land Registry Cymraeg – World's largest known transaction processing RDBMS >40 TB

Traditional Systems Landscape



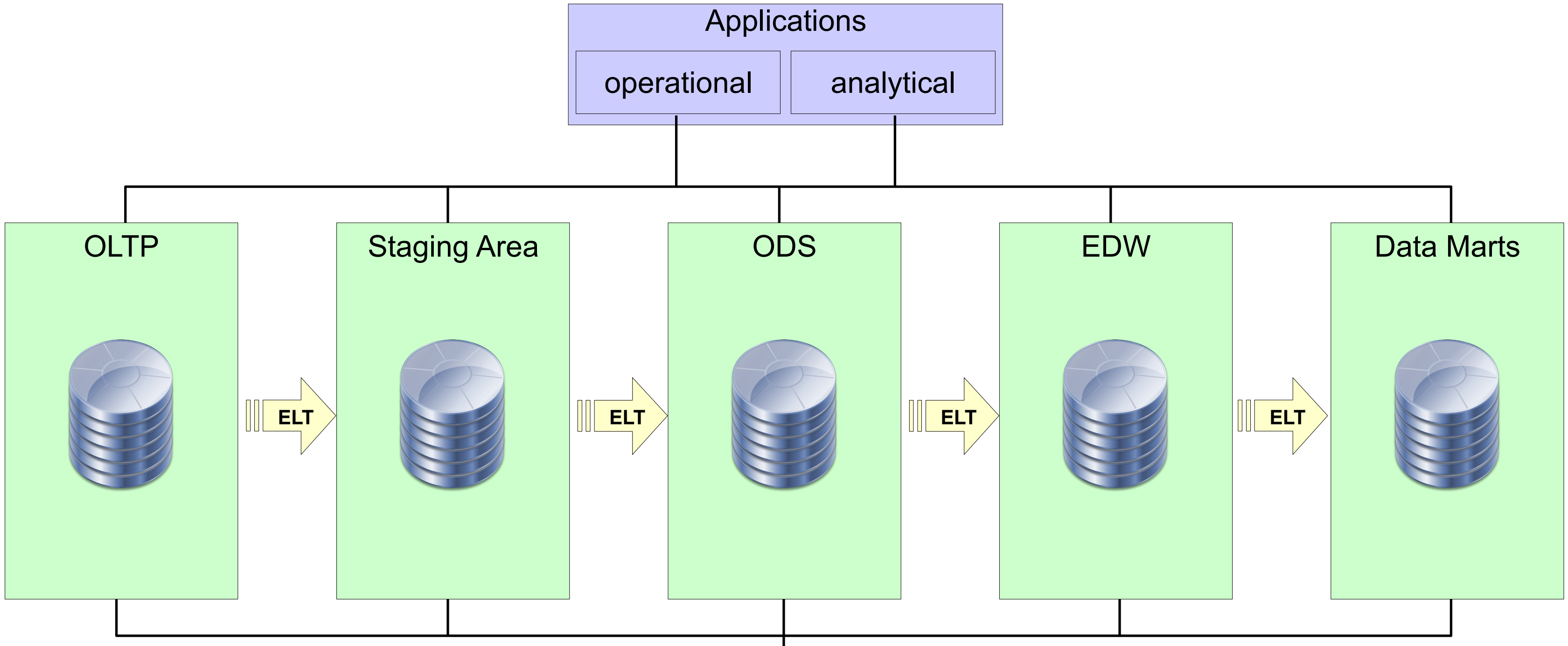
Historical reasons:

- Different access patterns
  - impact on performance
- EDW as the data integration hub
  - again, impact on performance
- Different life-cycle characteristics
  - and again, impact on performance
- Different Service Level Agreements (SLA)
  - Lack of broadly available workload management capabilities
  - Choice of lower cost-of-acquisition offerings

Negative ramifications:

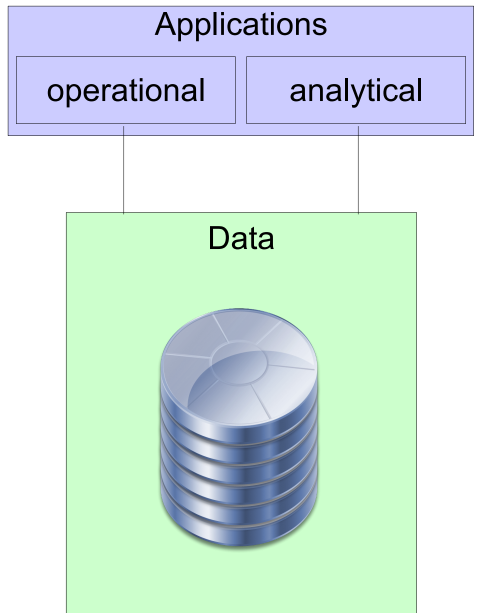
- Complexity
  - both in systems management and in applications
- Difficulties in supporting real time analytics
- Inability to match ever more demanding SLA requirements
- High total cost of ownership

Road to Visionary Systems Landscape



Benefits

- Uniform policies and procedures for security, HA, DR, monitoring, same tools, same skills, ...
- Efficient data movement within the system, often not involving network (ELT vs. ETL)
- Uniform access to any data for types of applications
- **Opportunity to remove, i.e. consolidate some of the layers, ultimately leading to a single database**



Benefits

- Uniform policies and procedures for security, HA, DR, monitoring, same tools, same skills, ...
- Efficient data movement within the system, often not involving network (ELT vs. ETL)
- Uniform access to any data for types of applications
- **Opportunity to remove, i.e. consolidate some of the layers, ultimately leading to a single database**

Challenges

- ✔ Mixed workload management capabilities
- ✔ Ensuring continuous availability, security and reliability
- ✔ Providing seamless scale-up and scale-out
- ⛔ Providing universal processing capabilities to deliver best performance for both transactional and analytical workloads **without the need for excessive tuning**

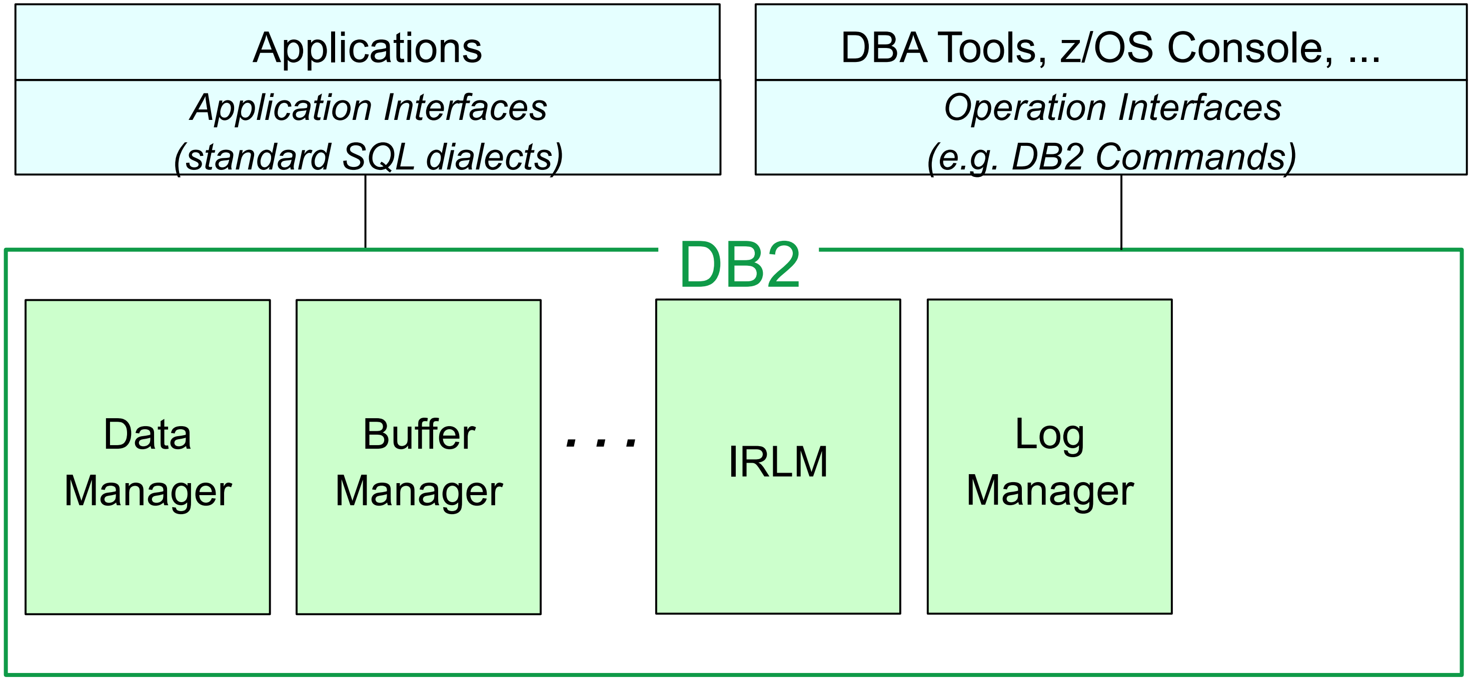
Approaches

- Columnar stores
- In-memory databases
- Hardware acceleration, special purpose processors
- Appliances

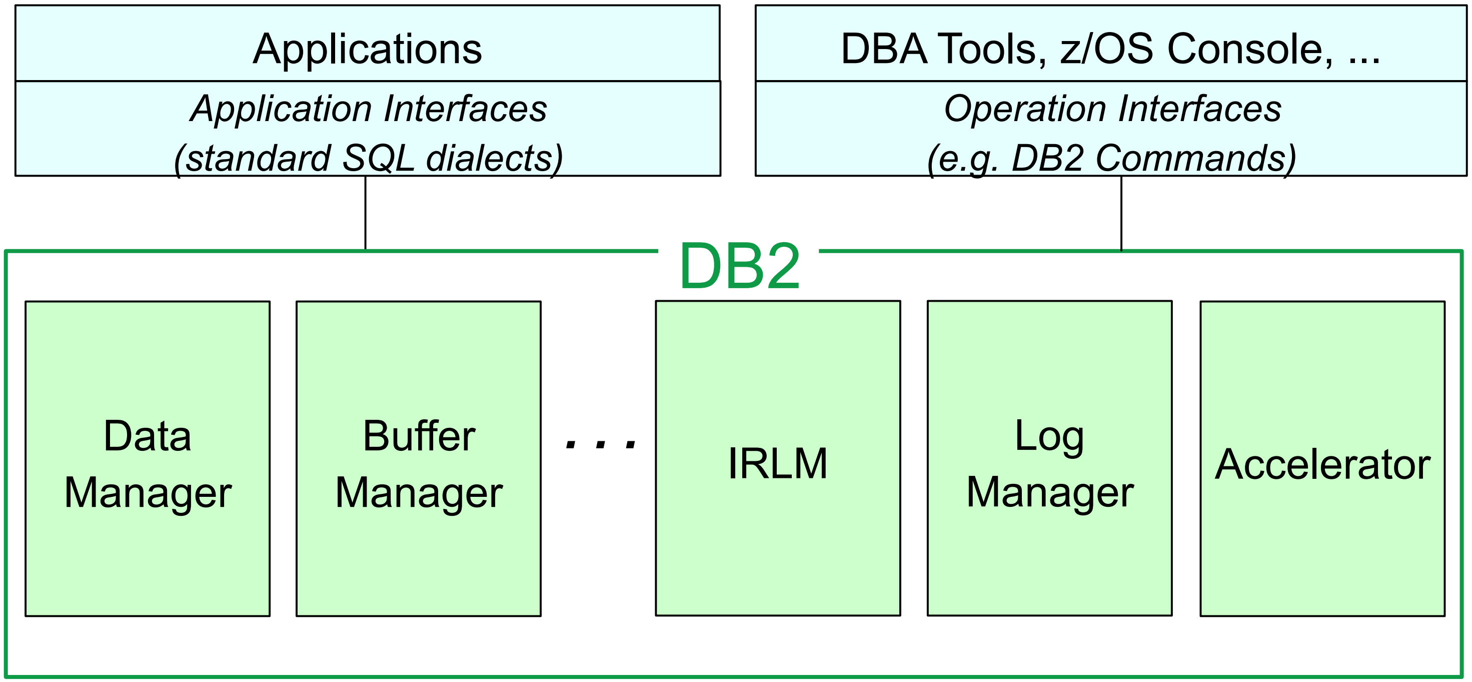
Building on proven technology base

- System z Data Sharing and Parallel Sysplex technology provides all the needed characteristics except one:
  - Special purpose processing for analytical workloads to minimize the need for manual tuning

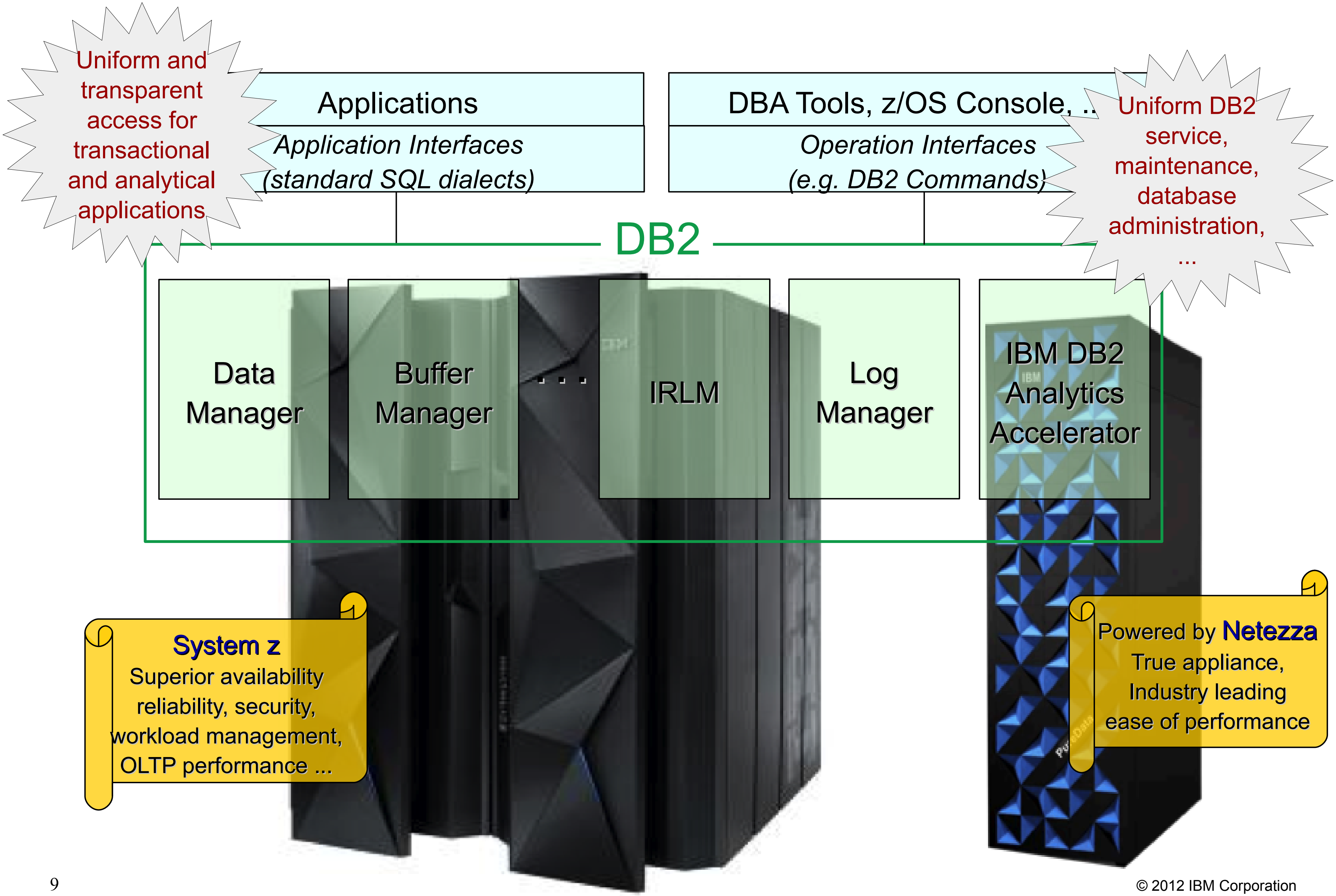
DB2 Components



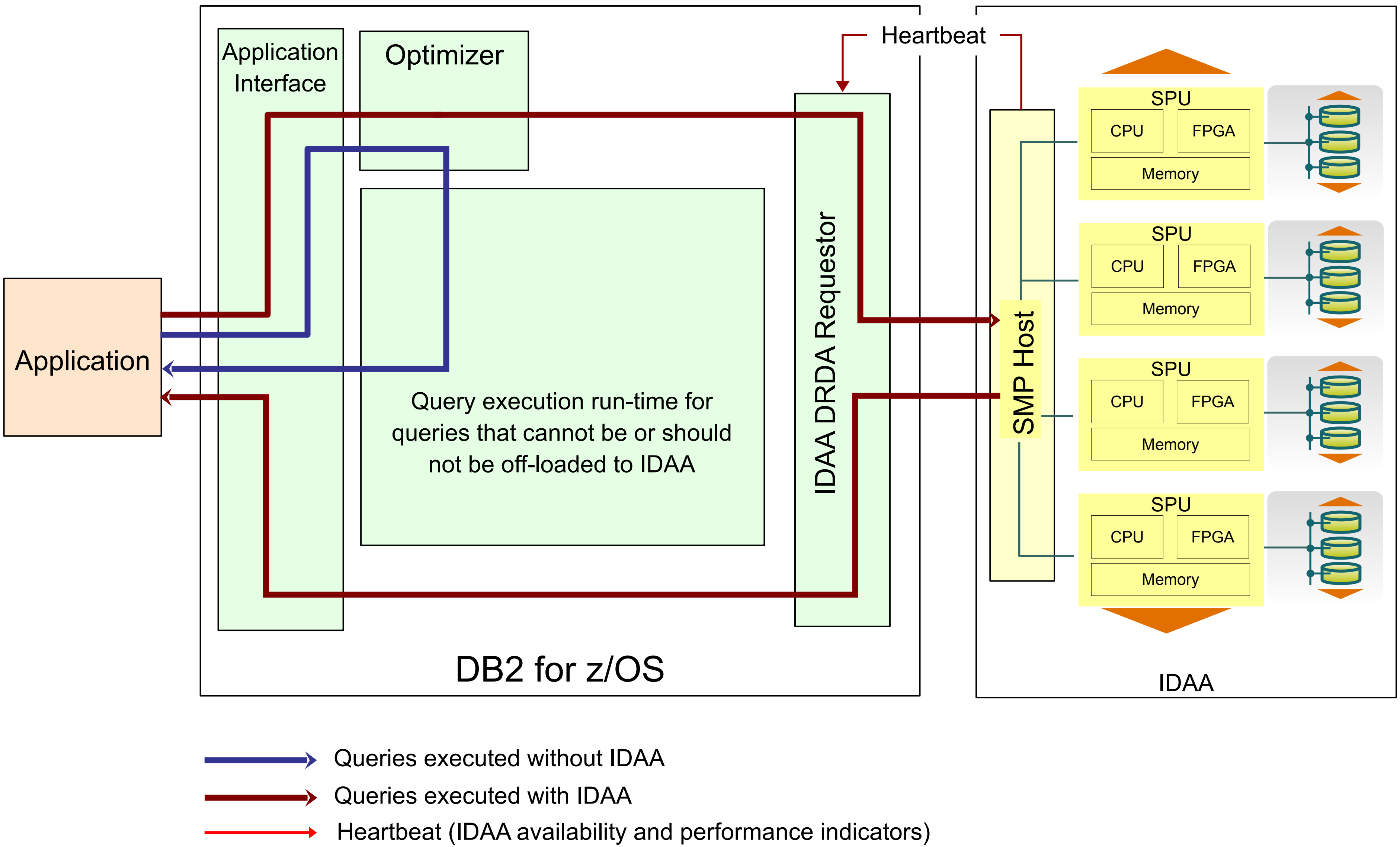
IBM DB2 Analytics Accelerator as a Virtual DB2 Component



DB2 Becomes a Hybrid Database Management System



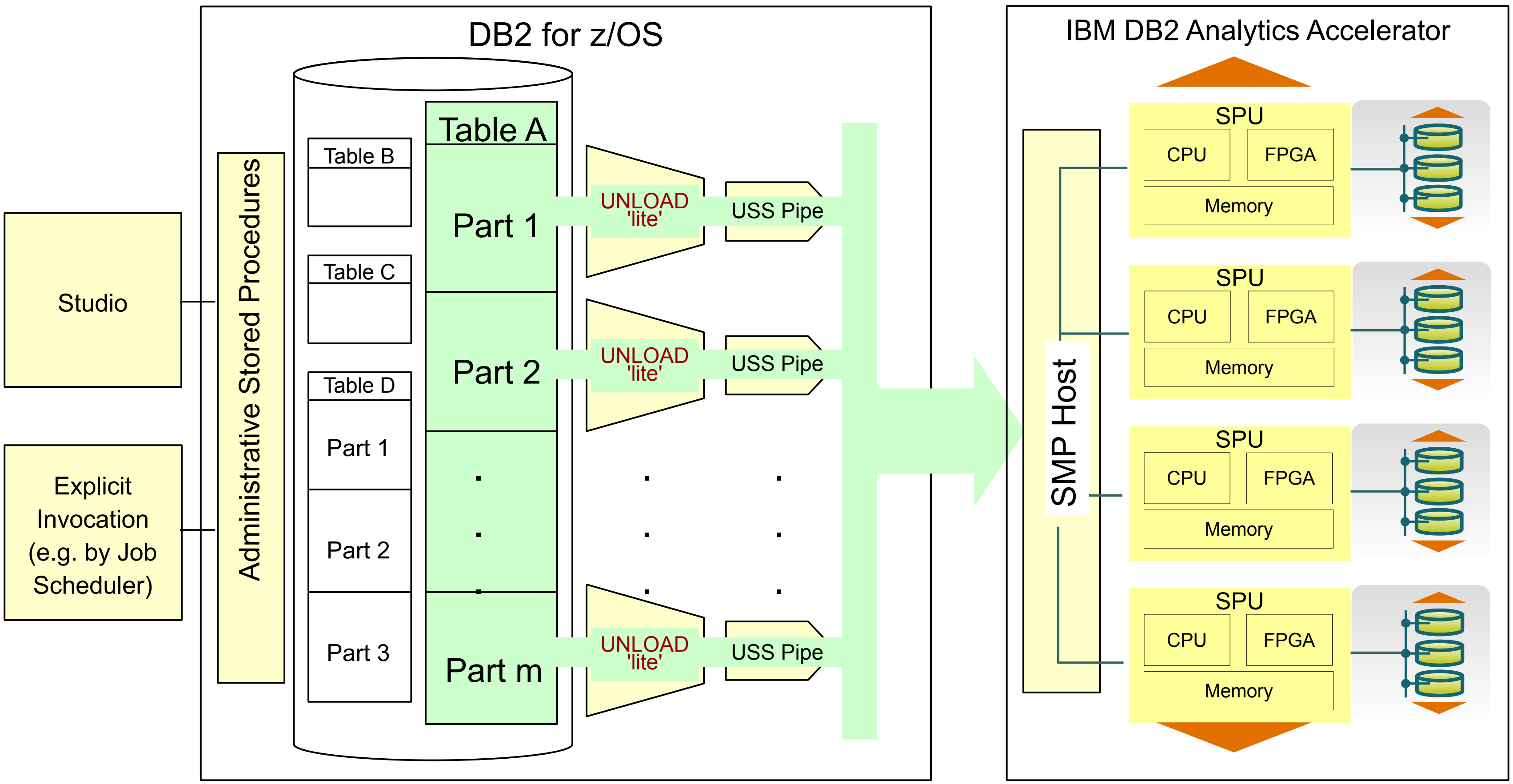
Query Execution Process Flow



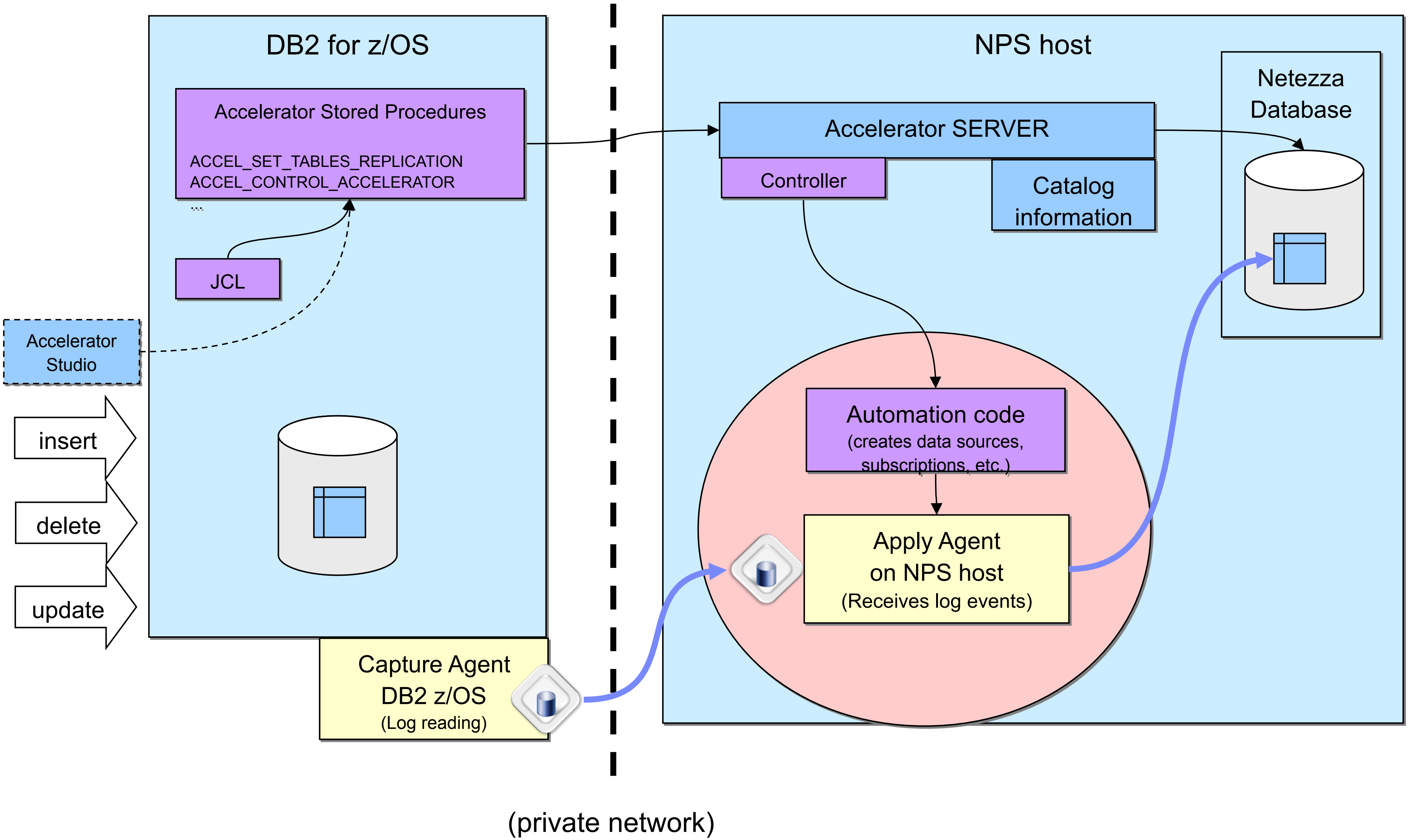
Synchronization Options

Synchronization options	Use cases, characteristics and requirements
Full table refresh The entire content of a database table is refreshed for accelerator processing	▪ Existing ETL process replaces entire table ▪ Multiple sources or complex transformations ▪ Smaller, un-partitioned tables ▪ Reporting based on consistent snapshot
Table partition refresh For a partitioned database table, selected partitions can be refreshed for accelerator processing	▪ Optimization for partitioned warehouse tables, typically appending changes "at the end" ▪ More efficient than full table refresh for larger tables ▪ Reporting based on consistent snapshot
Incremental update Log-based capturing of changes and propagation to the accelerator with low latency (typically few minutes)	▪ Scattered updates after "bulk" load ▪ Reporting on continuously updated data (e.g., an ODS), considering most recent changes ▪ More efficient for smaller updates than full table refresh

Initial Data Load and Refresh

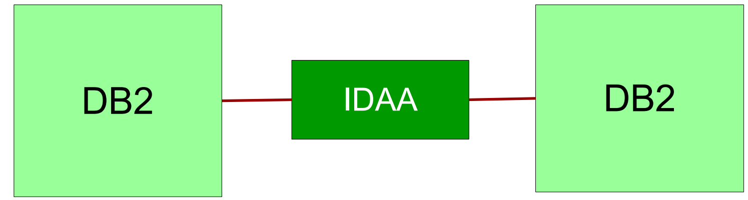


Incremental Updated Architecture

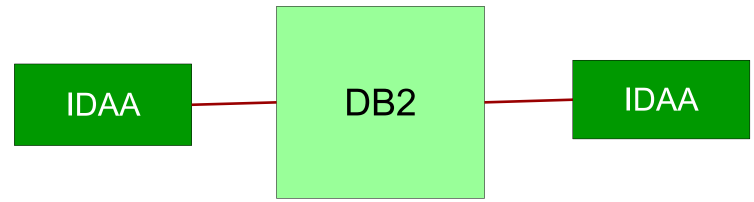


Connectivity Options

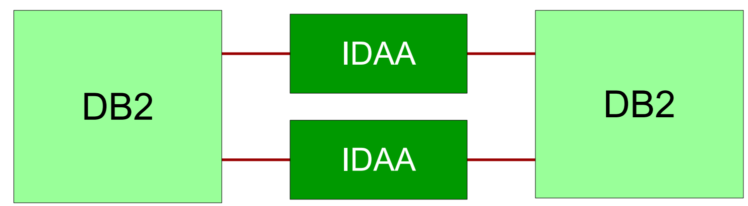
Multiple DB2 systems can connect to a single IDAA



A single DB2 system can connect to multiple IDAAs



Multiple DB2 systems can connect to multiple IDAAs



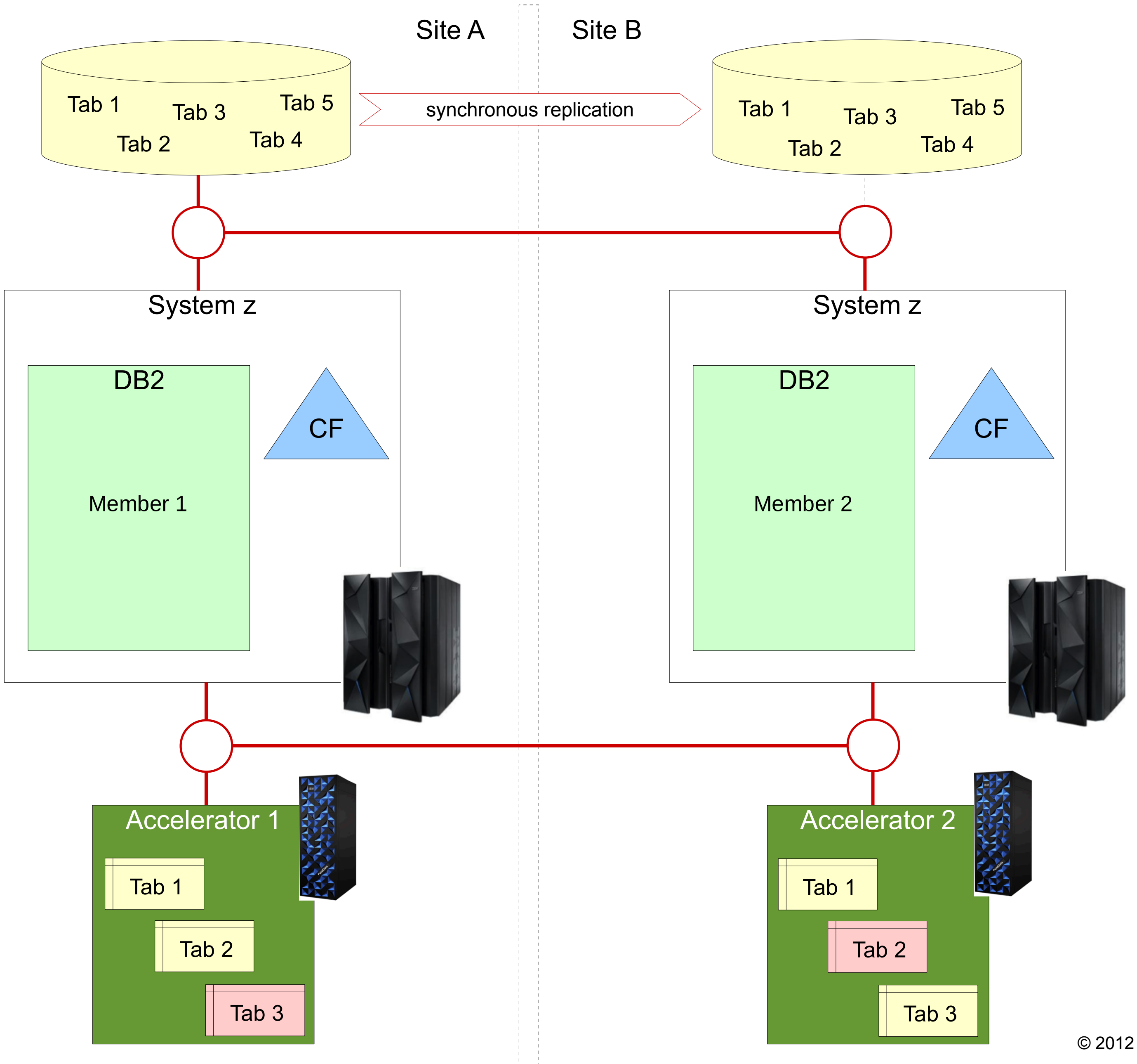
Better utilization of IDAA resources
Scalability
High availability

Full f exibility for DB2 systems:

- residing in the same LPAR
- residing in different LPARs
- residing in different CECs
- being independent (non-data sharing)
- belonging to the same data sharing group
- belonging to different data sharing groups

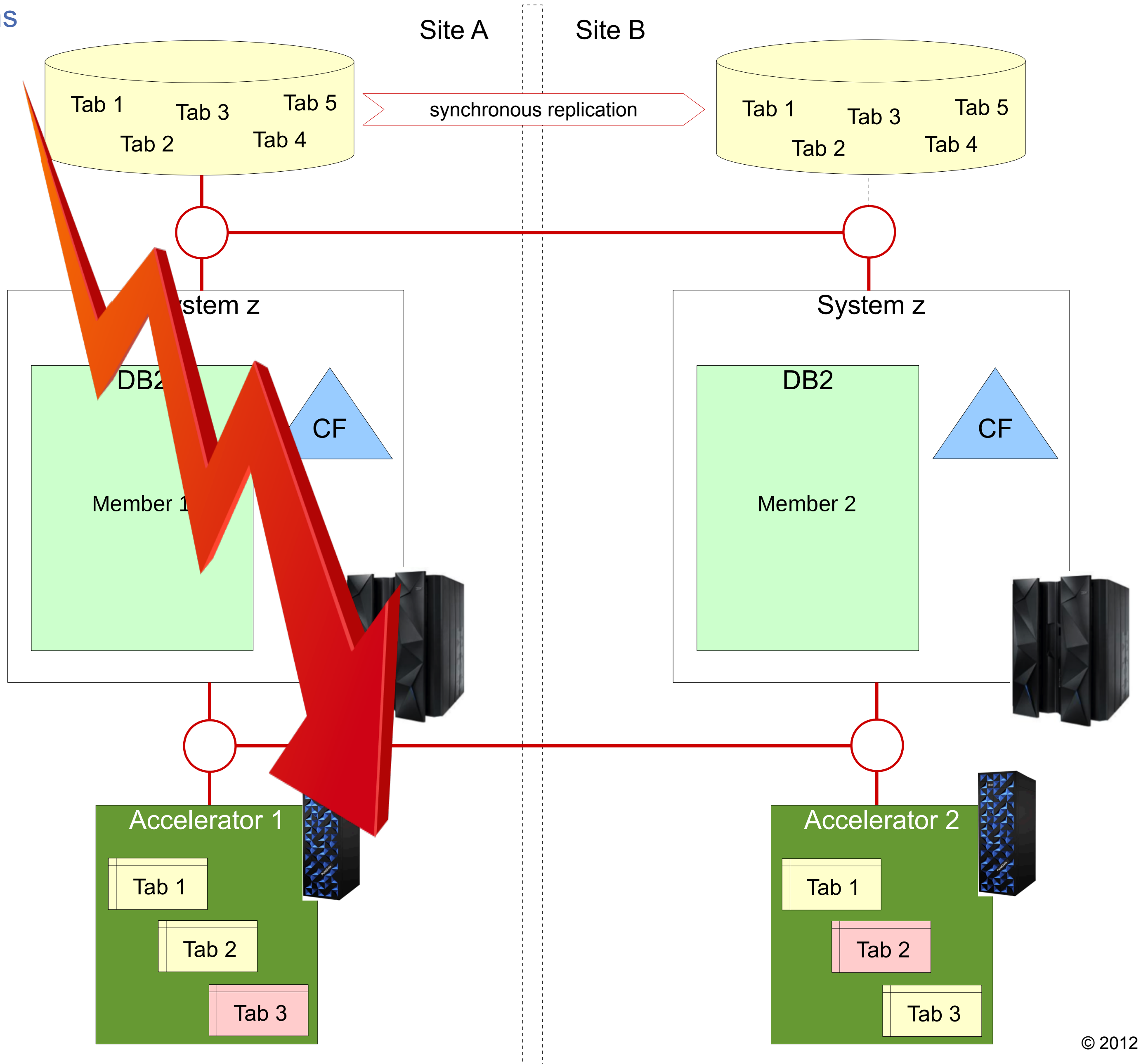
Disaster Recovery Configuration Example

Prior to disaster



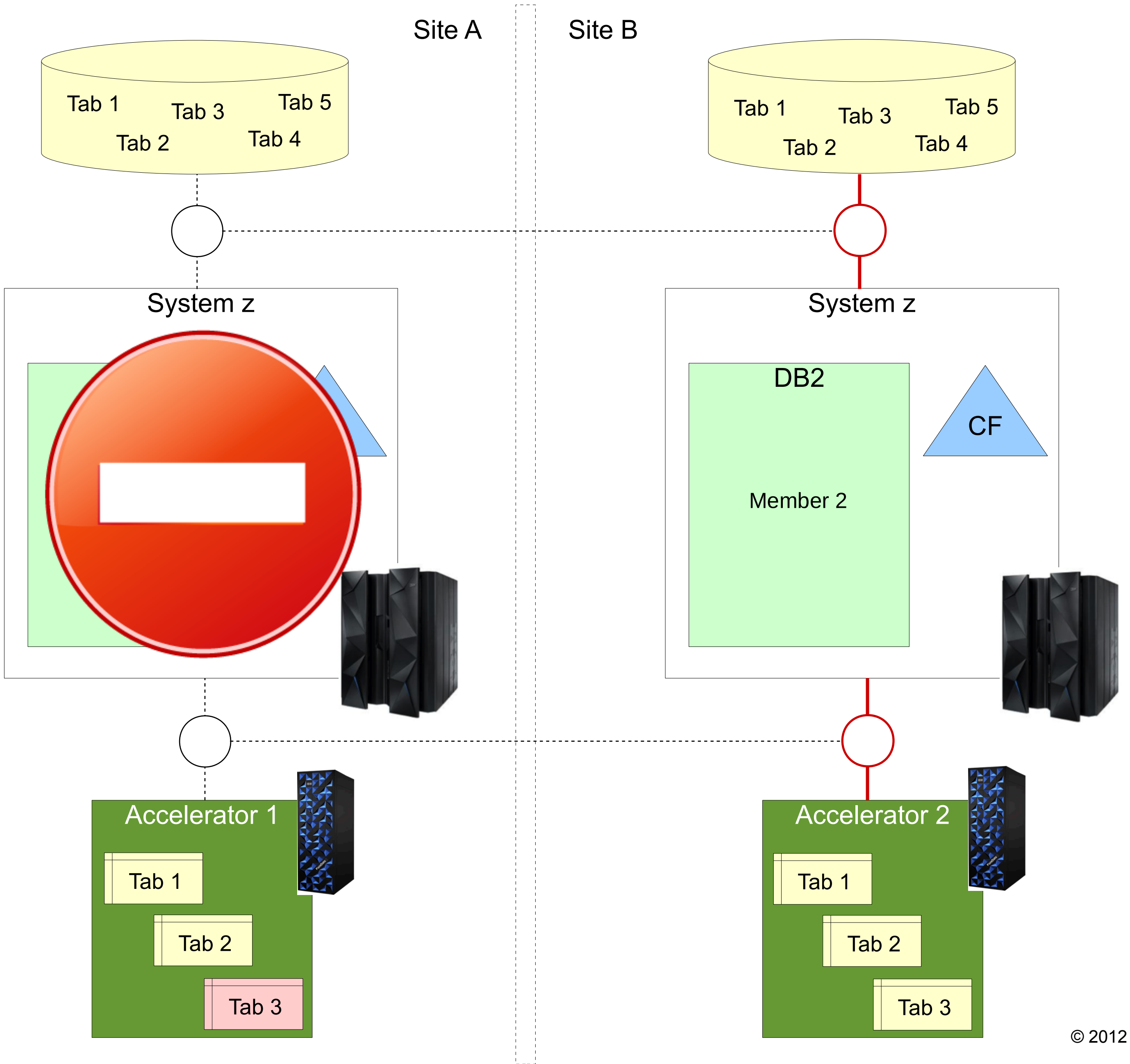
Disaster Recovery Configuration Example

Disaster happens



Disaster Recovery Configuration Example

After disaster



High Performance Storage Saver

- Most of the data in an ODS or EDW is static
  - The large tables are partitioned by time
  - Older partitions are never changed
  - The most recent partition is frequently changed
- Many DBMS vendors provide multi-temperature (MT) data solutions
  - MT concept is not the same as archiving, but they have lots in common
  - Level of sophistication in implementing MT varies
    - the industry leading solutions are so called 'near-line storage servers'
    - 'near-line' means 'near-online'
- MT's value proposition is twofold:
  - Move less frequently accessed data to cheaper storage
  - Improve performance for both queries and administrative operations accessing more recent data
- The drawback is degraded performance of queries that access old data
- Better solution is needed if the query access pattern includes both
  - Transactional, i.e. accessing limited amount of data, predominantly from the most recent partition, and
  - Analytical, i.e. accessing large amount of data across all the partitions
- IBM DB2 Analytics Accelerator can offer such a solution
  - High Performance Storage Saver
  - 'Online Storage Server' as opposed to 'Near-line Storage Server'
  - Netezza provides very large disk capacity at a fraction of cost of the System z disk subsystem
  - IBM DB2 Analytics Accelerator technology provides the basis for access to data irrespective of where they reside (on DB2 disks or Netezza disks)

Use Cases:

- Workload Isolation
  Ensure that the workload of one DB2 subsystem does not monopolize the resources of a shared accelerator. A development subsystem, attached to the same accelerator as a production subsystem, should not be able to drain all accelerator resources.

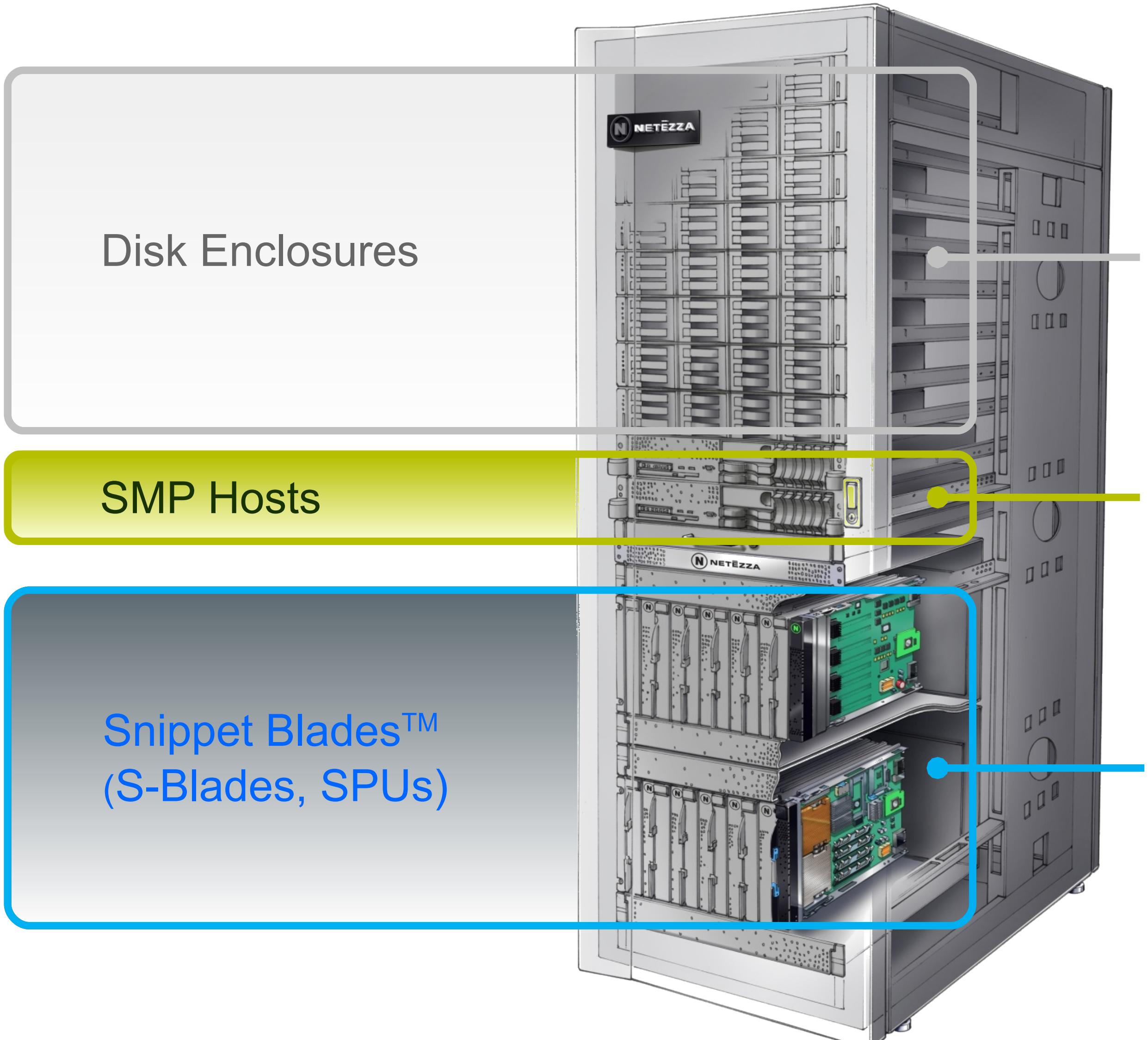
- Query Prioritization
  More important queries should be executed before and faster than less important queries that are sent from the same DB2 subsystem against the accelerator.

Special register CURRENT QUERY ACCELERATION

- Can be set implicitly by inheriting the value of the system parameter, or
- Explicitly by SET CURRENT QUERY ACCELERATION

Value	Description
NONE	No query is routed to the accelerator
ENABLE	A query is routed to the accelerator if it satisfies the acceleration criteria including the cost and heuristics criteria. Otherwise, it is executed in DB2. If there is an accelerator failure while running the query, or the accelerator returns an error, DB2 will return a negative SQL code to the application.
ENABLE WITH FAILBACK	A query is routed to the accelerator if it satisfies the acceleration criteria including the cost and heuristics criteria. Otherwise, it is executed in DB2. Under certain conditions the query will run on DB2 after it fails in the accelerator. In particular, any negative SQLCODE will cause a failback to DB2 during PREPARE or first OPEN. No failback is possible after a successful OPEN of a query.
ELIGIBLE	A query is routed to the accelerator if it satisfies the acceleration criteria irrespective of the cost and heuristics criteria. Otherwise, it is executed in DB2.
ALL	A query is routed to the accelerator. If it cannot be executed there, the query fails and a negative return code is passed back to the application.

Powered by Netezza

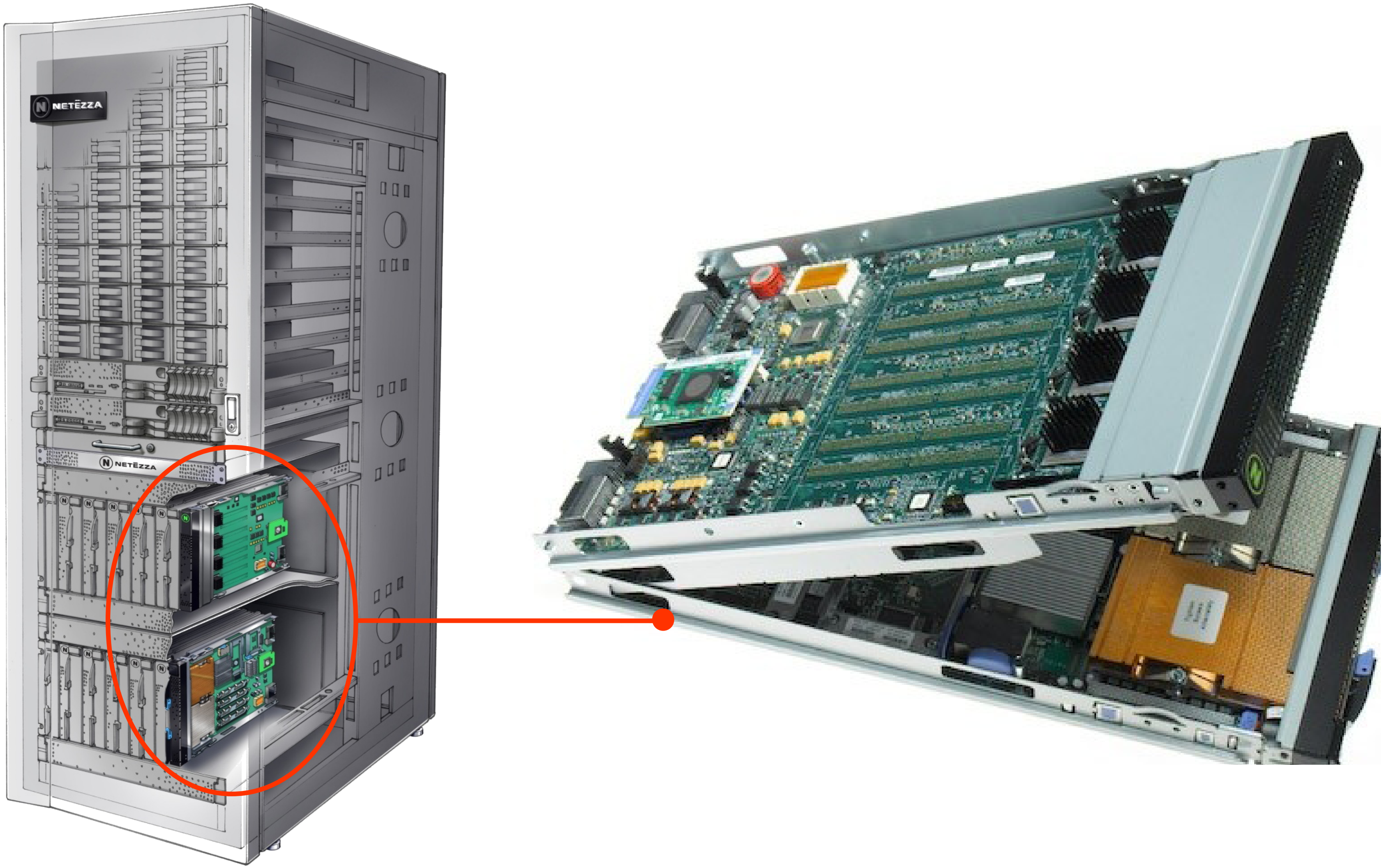


Slice of User Data
Swap and Mirror partitions
High speed data streaming
High compression rate
EXP3000 JBOD Enclosures
12 x 3.5" 1TB, 7200RPM, SAS (3Gb/s)
max 116MB/s (200-500MB/s compressed data)
e.g. in model 1000-12:
8 enclosures → 96 HDDs
32TB uncompressed user data (→ 128TB)

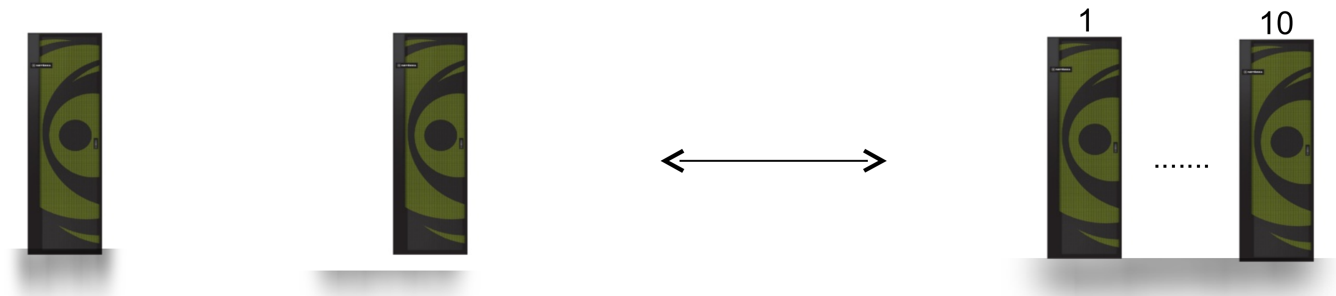
IDAA Server
SQL Compiler, Query Plan, Optimize
Administration
2 front/end hosts, IBM 3650M3 or 3850X5
clustered active-passive
2 Nehalem-EP Quad-core 2.4GHz per host

Processor &
streaming DB logic
High-performance database
engine streaming joins,
aggregations, sorts, etc.
e.g. in model 1000-12: 12 back/end SPUs
(more details on following charts)

The Netezza S-Blade™



Scaling Out



Model	1000-3	1000-6	1000-12	1000-24	1000-36	1000-48	1000-72	1000-96	1000-120
Cabinets	1/4	1/2	1	2	3	4	6	8	10
Processing Units	24	48	96	192	288	384	576	768	960
Capacity (TB)	8	16	32	64	96	128	192	256	320
Effective Capacity (TB) *	32	64	128	256	384	512	768	1024	1280

Capacity = User data space
Effective Capacity = User data space with compression (4x compression assumed)