Netezza

Professor Li-Yan Yuan
University of Alberta, Canada

- Big Data and data-warehouse
- Source: IBM Netezza White-paper

What Is Netezza

Netezza designs and markets

- high-performance data warehouse appliances and
- advanced analytics applications

for uses including

- enterprise data warehousing,
- business intelligence,
- predictive analytics and
- business continuity planning.

Why Netezza?

Data volumes are so huge, and the need for business-critical information so sweeping, that this creates challenges for organizations across many industries:

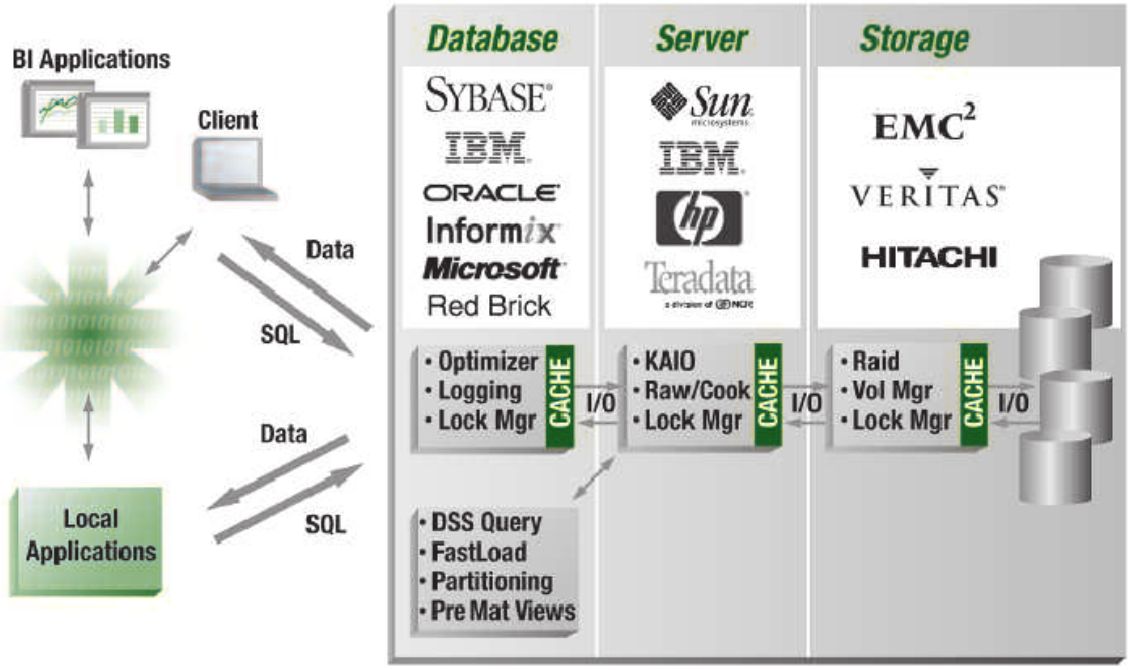
- For the major wireless telecommunications carrier:
  - 250 million call detail records daily
- For the health-care provider
  - sophisticated data mining of terabytes of operational and patient data
- For the large online retailer
  - recording every click of millions of customers shopping habits,
  - analyzing billions of rows of data in order to develop targeted promotions.
- For the grocery retailer,
  - perform complex market basket analysts against detailed line-item level transactional data

current data warehouse systems are based on older architectures that weren't designed to handle todays demands for querying enormous amounts of data.

Traditional Data Warehouse Systems

The High Cost of General-Purpose Solutions

- Server
- Storage
- Software



Barriers to Performance

- General-purpose servers:

  – These are the same computers used in data centers as web servers, email servers or application servers.

  – Data mining processing can involve extremely large sets of data, and query requirements are quite different.

- General-purpose storage:

  Most general-purpose storage arrays require time-consuming, careful synchronization of loaders and data striping mechanisms to ensure that data is distributed so that it can be accessed efficiently by business intelligence users.

  Finding the specialized expertise to properly configure the storage system usually means engaging a costly professional services firm.

- General-purpose database:

  Obtaining maximum performance from a data warehouse requires a close marriage between its software and hardware architectures.

  The full power of general-purpose database management systems (DBMS) such as DB2 or Oracle is lost because they are designed to extract optimal performance out of even the most advanced servers and storage.

Barriers to Efficiency

The sheer inefficiency of patchwork solutions creates cost, complexity and waste:

- Inefficient use of administrators time:

  – How to configuring storage devices from EMC (for example), servers from HP or IBM, and database management software from Oracle for the demands of terascale query processing

- Inefficient installation:

- Inefficient system management:

  Patchwork solutions become increasingly difficult to manage as core products evolve, especially as vendors upgrade their offerings at different times.

- Inefficient data flow:

  – Query processing on a general-purpose system is extremely cumbersome,

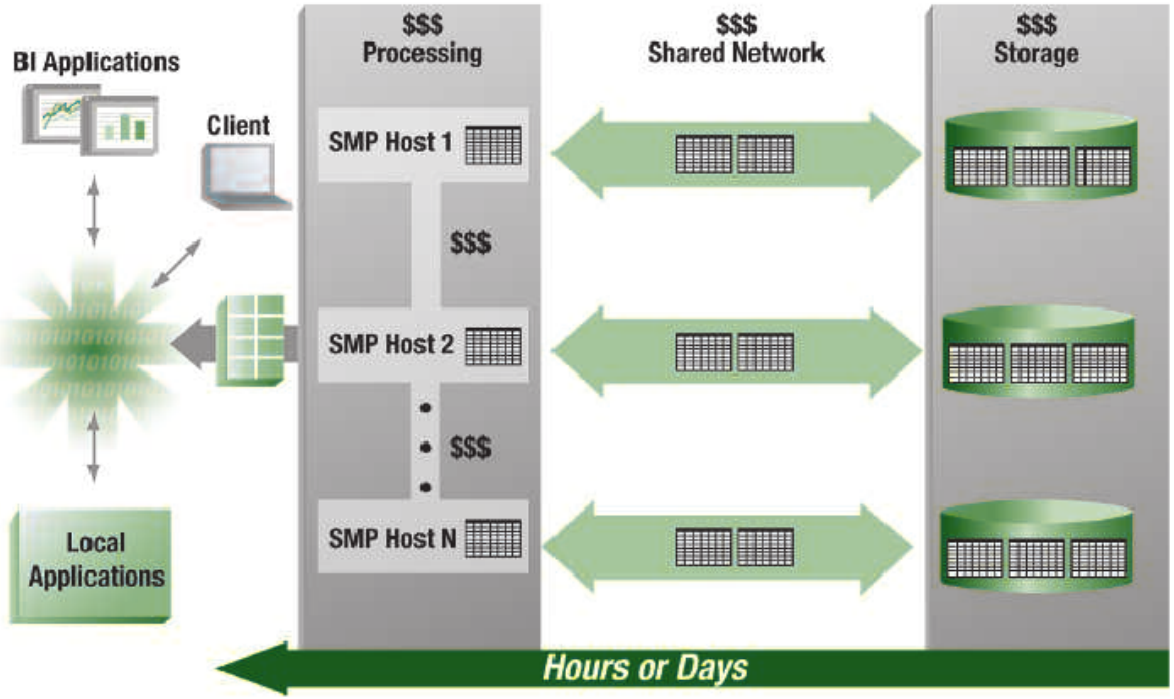
Inability to Scale

- Data volumes:
- Complexity of queries:
- Real-time response:
- Number of users:

Traditional Data Flow

Unlike OLTP, data warehousing is all about data shuffling:

– moving large quantities of data through the systems analysis and processing engine as efficiently as possible



Traditional Multiprocessing Architectures

- Symmetrical Multiprocessing (SMP)

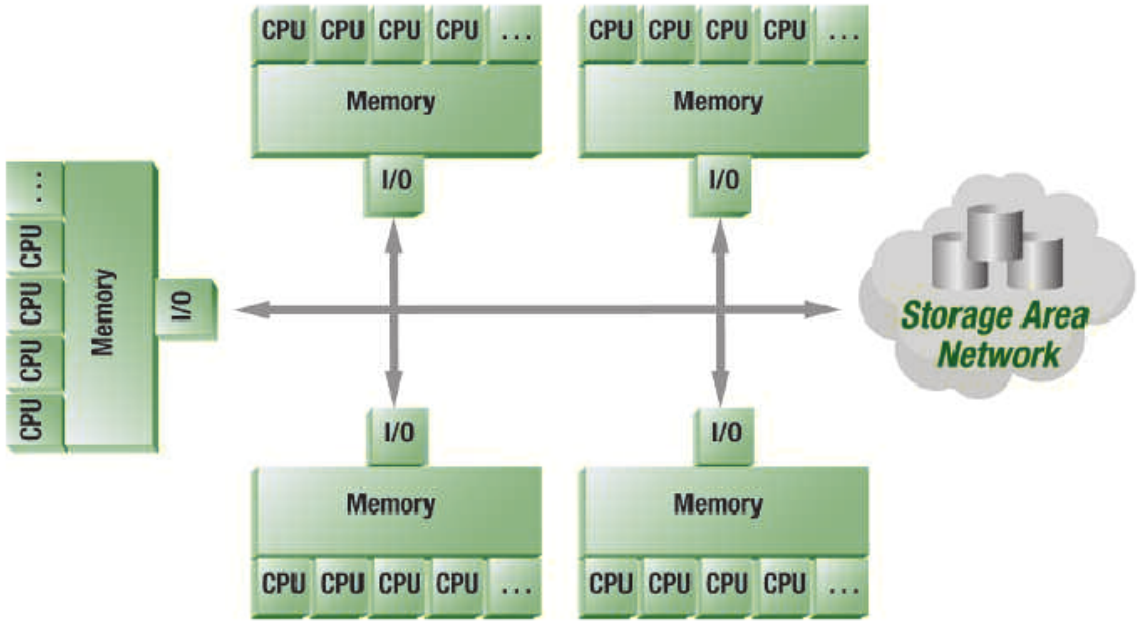
  SMP systems consist of several processors, each with its own memory cache.

  The processors constitute a pool of computation resources, on which threads of code are automatically distributed by the operating system for execution.

- Massively Parallel Processing (MPP)

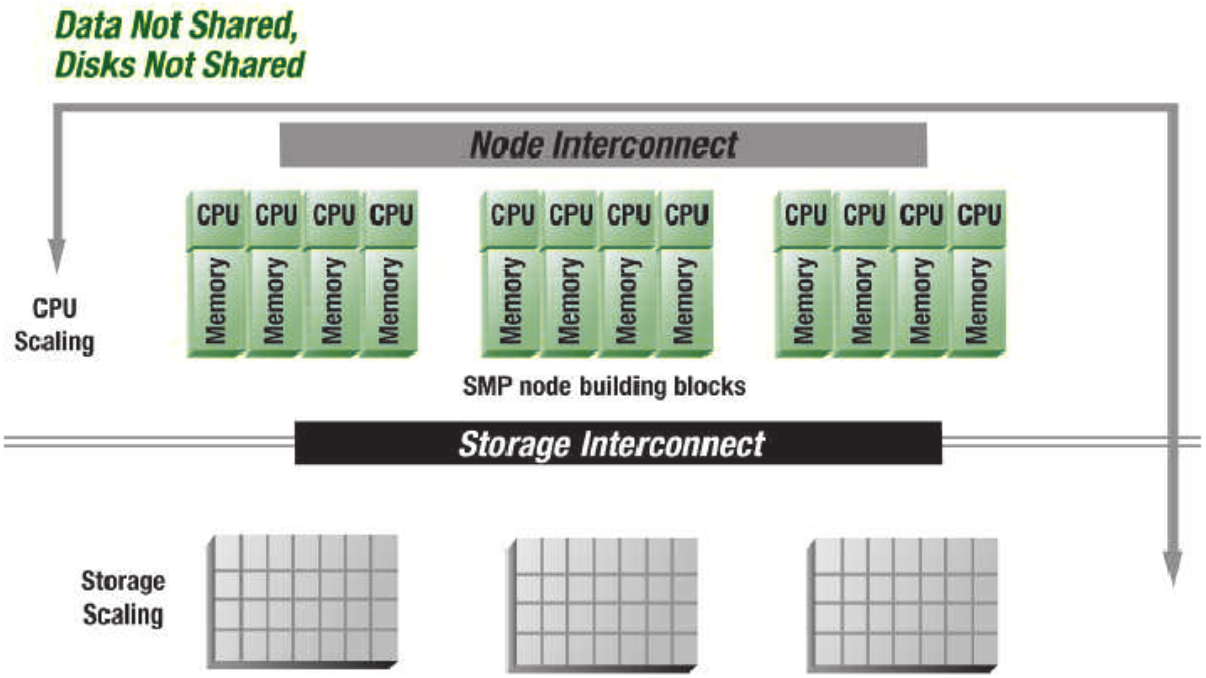
  MPP systems consist of very large numbers of processors that are loosely coupled. Each processor has its own memory, backplane and storage, and runs its own operating system.

MPP on Clustered SMP



Mainstream Examples of Traditional Architectures

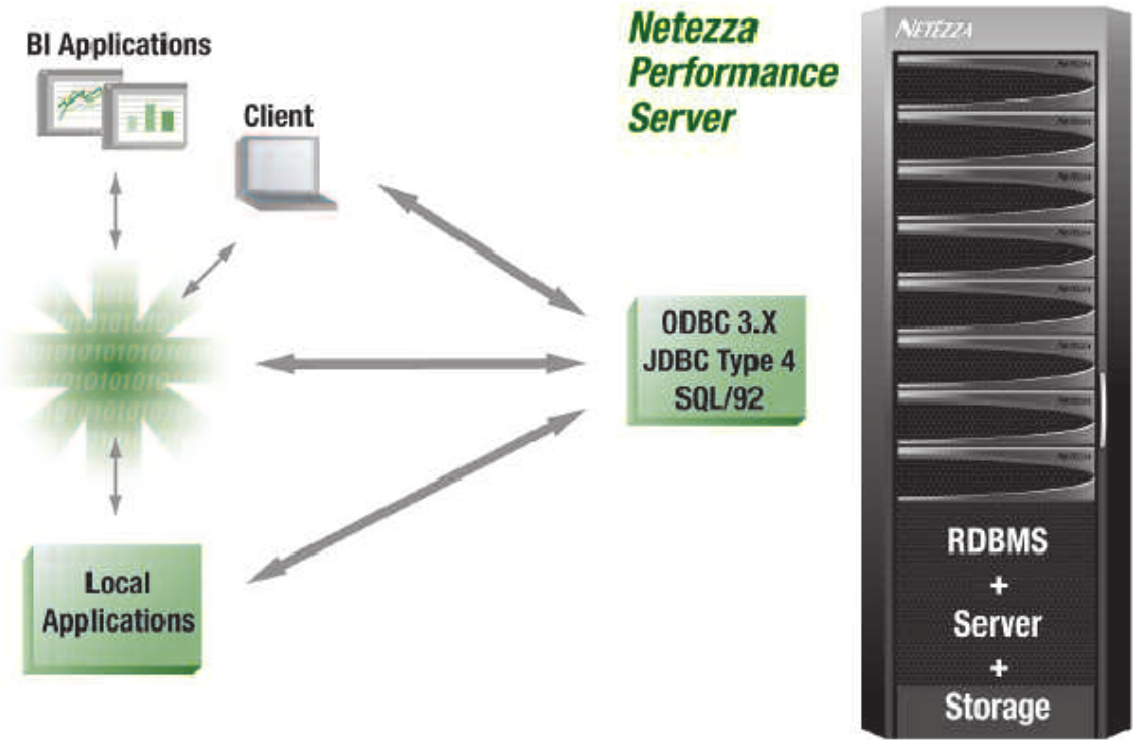
Shared Nothing MPP (Such as Teradata)



Netezzas Data Warehouse Appliance

Performance, Value, Simplicity

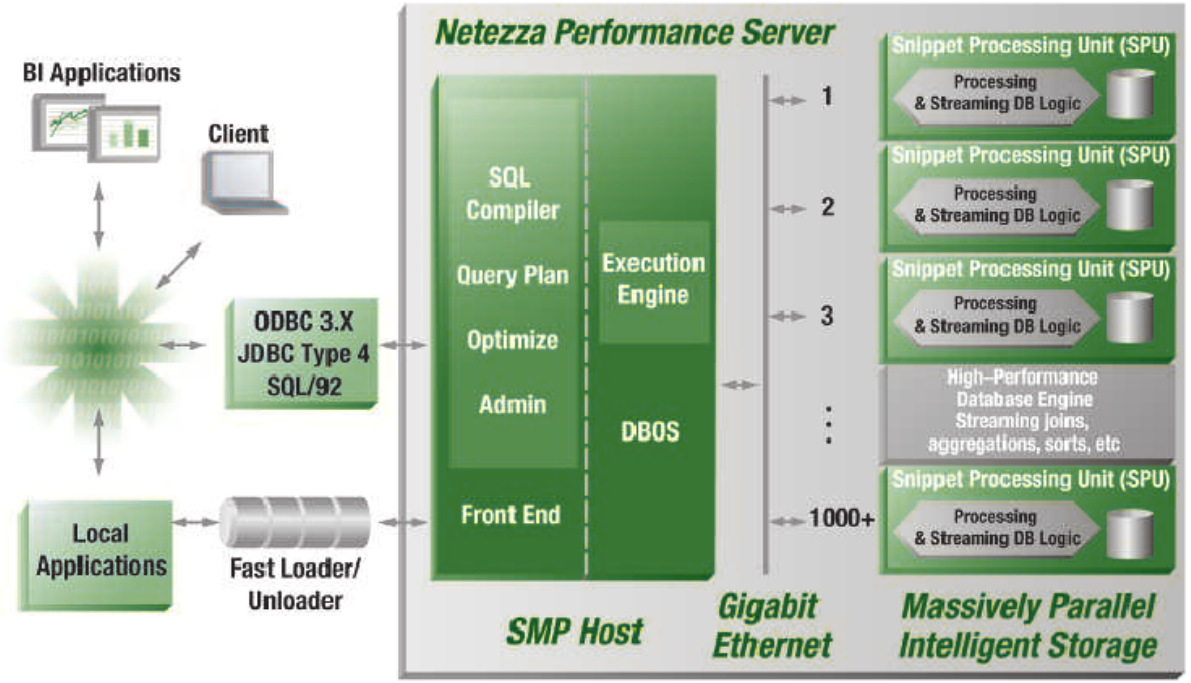
- A fully integrated device built for a single purpose: to enable real-time business intelligence and analytics on terabytes of data.
- A single scalable platform based on open standards and commodity components



Data Flow - The Netezza Way

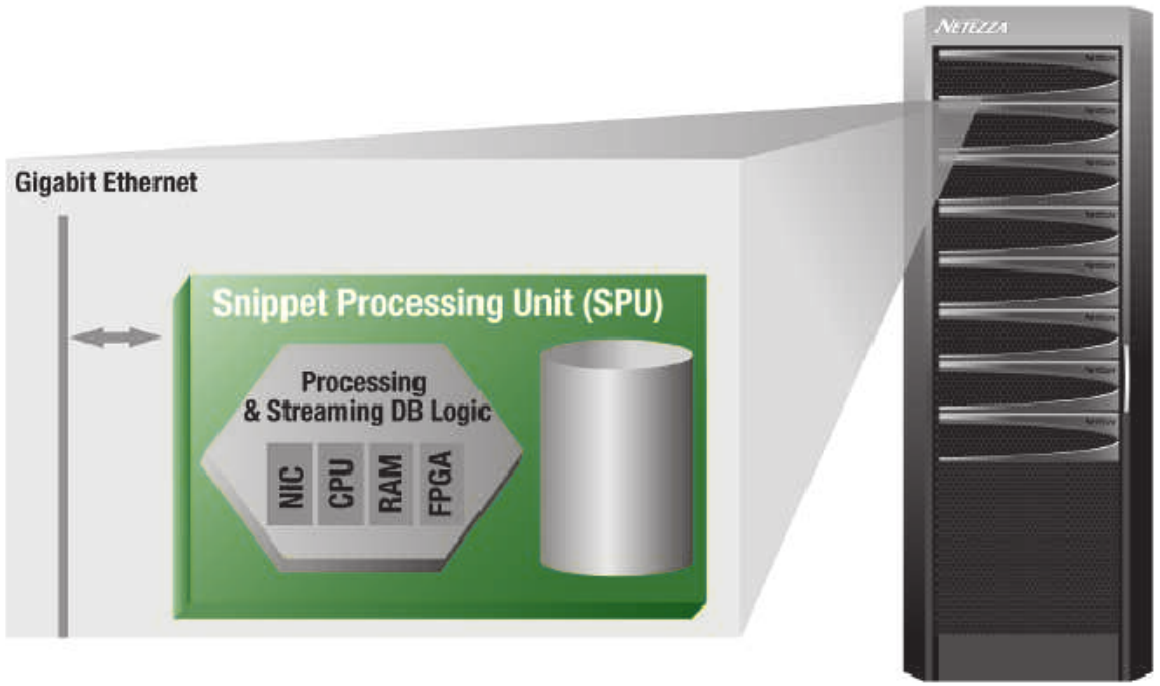
The architecture of the NPS appliance is built upon two guiding principles:

- Performance and scalability goals can be met using elements of both SMP and MPP,
- Moving processing intelligence to a record stream adjacent to storage



The Snippet Processing Unit (SPU)

Each SPU is an intelligent query processing and storage node, and consists of a powerful commodity processor, dedicated memory, a disk drive and a field-programmable disk controller with hard-wired logic to manage data flows and process queries at the disk level



Netezzas AMPP architecture

A two-tiered system designed to handle very large queries from multiple users.

- The first tier is a high-performance Linux SMP host.

  – The host compiles queries received from BI applications, and generates query execution plans. It then divides a query into a sequence of sub-tasks, or snippets, that can be executed in parallel, and distributes the snippets to the second tier for execution.

- The second tier consists of dozens to hundreds or thousands of Snippet Processing Units (SPUs) operating in parallel.

Key Differences and the Netezza Advantage

- Data flow
- Storage Connection
- Degree of Integration
- Linear Scalability

Interfaces

- SQL
- ODBC,JDBC
- Java, C++,Python
- Hadoop, MapReduce

Performance

- Load Time: 2TB/hour
- Backup/Restore: 4TB/hour