The University of Phoenix Wins Big with SAS® Grid Computing

Daqing Zhao, PhD, VP Analytics
Aptimus, Apollo Group
March 24, 2009
Business Context

- **Apollo’s Univ. of Phoenix**
  - Largest private university with 360K students worldwide
  - Largest online student population (80%)
  - Over 200 campuses

- **Associates, Bachelors, Masters and PhDs, as well as High School programs**
  - US has 18 million college students and 70 million US adult population don’t have college degree

- **Aptimus is marketing arm of Apollo edu institutions**
  - Large number of student leads per month and 30K enrollments
  - Large amount of data, high economic impact
  - Need sophisticated analyses, and scalable data processing and statistical computing power
CRM Strategy

- Every student can have own optimal learning environment
- Right message to the right person at the right time
- Marketing and recruiting, relevant message to recruit only people who are truly interested in education
- Student segmentation for services, and predict retention
- Identify risk factors, goals and objectives
- Improve quality of education and student experience and graduation rate
Data and Analytics Strategy

- Student and faculty event level data
  - Banner and search behavior
  - Web site behavior
  - Email behavior
  - Demographics and psychographics
  - Call center behavior
  - Classroom behavior
  - Surveys and web logs
  - Experimental designs

- For Internet media players, Yahoo, Google, MySpace, etc., the game is to optimize from impression to click to lead form submission

- For University of Phoenix, we manage and optimize not only these, but also lead to enrollment conversion, student educational experience, student retentions

- Large data problems
Grid computing

- The last couple of years, Moore’s Law has flatten off
  - Memory bandwidth is the bottleneck for processing power of a single CPU
  - Not clock speed
- Parallel computing
  - Multiple processors access the same copy of memory
  - Expensive
  - Many SAS installations are still on these big servers
- Distributed or grid computing
  - Multiple CPUs on multiple computers, have own memory
  - High Ethernet bandwidth
  - COTS computers (consumer off the shelf)
- Only grid computing guarantees to scale
  - SAS some procs are multithreaded, others are not
  - For large datasets, disk IO is the bottleneck for processing
- SAS Grid can be deployed on a cluster of inexpensive servers
Sun Hardware for SAS Grid

- Four identical low cost Sun X4600 servers, four processors each, duo core AMD 3.0GHz processors, 64GB RAM per server
  - Solaris 10
  - Total 32 nodes
  - Quad Gigabit Ethernet

- 40TB of SAN clustered file system, ST9990V
  - Some 146GB disks, two fiber channel controllers, 30 MB/sec disk IO throughput per core, no backups, RAID 5
    - SAS temp work space
    - Project directories and SAS install directory
  - Some 300GB disks that are backed up, RAID 5

- Grid Manager and Meta Data server on Sun T5220 machine

- All servers attached to the shared 40TB SAN disk
SAS Installs and Usage Scenarios

- Single SAS installation on the shared disk
  - All slaves read from the same SAS install
  - No need to install SAS for new servers
  - Grid manager and meta data server

- Analyst can log in each server or servers independently for smaller tasks
  - All servers are interchangeable
  - Each server is quite powerful

- All servers can work together as a grid on larger tasks
  - With simple modifications of code, no need to change code as data scale
  - Every server can access the grid
  - SAS Grid Manager manages jobs on the nodes

- SAS EM and EG connect to the grid and submit jobs
SAS Setup at University of Phoenix

- Servers reside at the same location as the data warehouse
  - No need to transfer large amount of data over the Internet

- We have all SAS modules installed as fat clients, with all the SAS functions available on PCs
  - Smaller data sets can be sliced and diced locally
  - Development can be done on Windows platform

- System is scalable and fault tolerant, reduce IT SLA requirement
  - If one or two servers are down, we can still process data, just slower
  - Meta data and grid manager server also has a fail over server
  - We can increase computing power by simply adding additional servers
    - No need to upgrade entire server system
    - We can add newer model machines to work with existing nodes
    - No need even to install SAS on new machines
Example: Web media optimization

- Web log data for media optimization
- Optimize clicks and application and enrollment yield, at CPM, CPC, and CPL
- Optimize by exposure, session, user behavior and targeted messages
- Data are at user impression level information, based on anonymous cookie identifier
  - For example, one website has 25GB data per day, or 750GB per month
  - Longitudinal data sets with data appends and feature generation in the TB range
  - Low signal from large amount of data
- Disk IO is bottleneck in processing speed
Grid Solution

- Using SAS Base, SAS Connect, and SAS Grid Manager
  - Allow each node to process one day of data, and reading and writing compressed files
  - Aggregate and process user level data and take appropriate samples
  - SAN disk supports high disk IO speed

- Build predictive models using SAS/stat and SAS Enterprise Miner
  - CPU intensive
  - Multiple models on the same data set to optimize model performance
  - Many models for different targets at the same time

- With sufficient computing power, we can add:
  - Complexity of targeting and optimization
  - Update models more frequently
  - Search large set of predictors
  - Get responsive, interactive queries on large data sets
Web log processing example

- Single day log file processing takes 30 minutes (14MB/sec) CPU time
  - On single SAS session, it takes 15 hours to process 30 days

- Using 15 nodes on a grid, parallel read and write
  - 30 minutes to process 15 days
  - 1 hour to process 30 days, and 1 hour to aggregate (200MB/sec)
  - Total 2 hours to process one month of data

- Sorting, 8 hours CPU time, but 5 hours real time
  - Cannot utilize grid directly
  - PROC SORT is multithreaded
  - Total 7 hours processing time is enough for us

- Alternative, more elaborate strategy is to split each day file based on sorting key
  - Aggregate by buckets of keys and sort separately
  - Aggregate all at the end
Grid search for best models
Why Sun/Windows for server/client

- Sun consultants are part of the team, providing great services from design to implementation
- Solaris 10 a mature operating system
- OS support for fast disk IO
- SAS on Windows with GUI to simplify development efforts and achieve high productivity
- Complete development tools for additional usage of the servers, so that we don’t compromise flexibility for other available code, C++, Perl, R, etc.
- Large number of open source software available on Solaris, if needed
- Low cost high power Sun servers readily available
- Blazing fast, rock solid, very responsive, and easy to use
- Favored by system administrators as well as end users
Why we chose SAS

- SAS consultants are part of the team, providing great services from design to implementation
- SAS is the most tested software, same scripts run on different platforms, 30+ years in the market, widely used by drug companies, financial institutions, and in academic research
- Powerful SAS base, stat, OR, ETS, allow us to transform data, sample data, and generate reports efficiently and easily
- SAS’s efficient use of work temp disk space as well as RAM allows processing of extremely large data sets
- SAS ACCESS connect to Oracle and MySQL databases easily
- SAS EG allow easy access of SAS by analysts
- SAS EM power model building tools, grid enabled, with convenient GUI
- SAS Text Miner for web page classifications
- SAS Grid Manager schedules tasks optimally
- Great SAS user community, excellent SAS tech support and training
Conclusion

- What we have is not ideal, not “share nothing”
- But we can implement many similar strategies to scale
  - Parallel read/write/compute
  - Also benefit from multithreads
- The setup provides computation horsepower for our data size and complexity of analysis
- Scalability and fault tolerance
- Flexibility of conducting analysis using SAS
- Cost and performance
- High computing power make complex queries and analysis more or less interactive
- We evaluated TeraData, and will continue to, but what we have seems sufficient for now