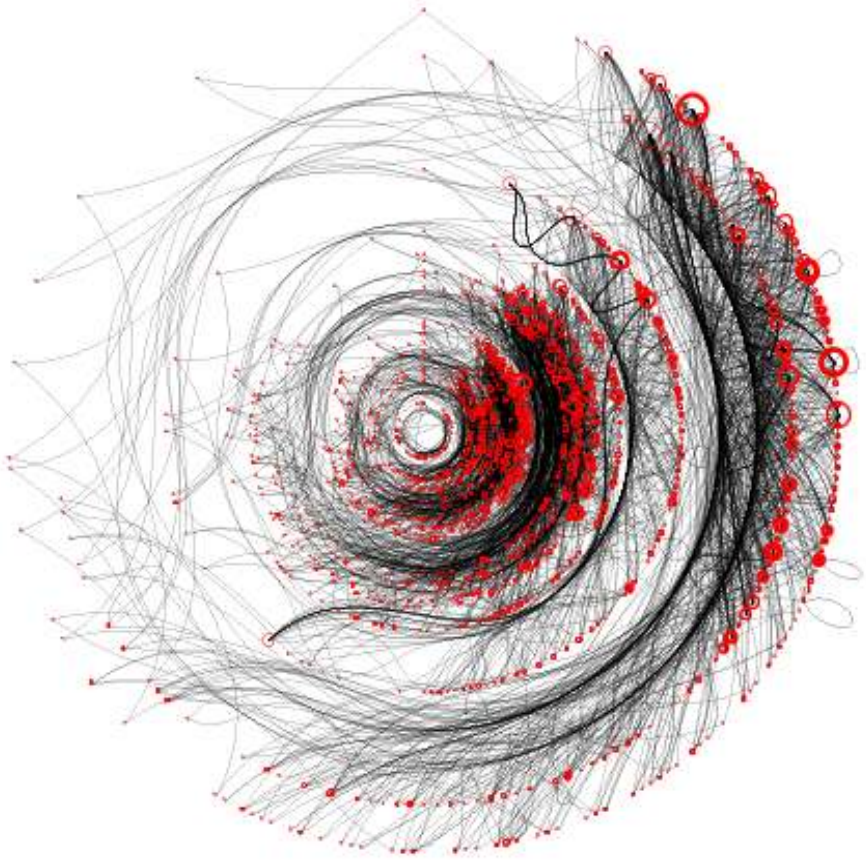


LEVERAGING HPC AND ENTERPRISE ARCHITECTURES FOR LARGE SCALE INLINE TRANSACTIONAL ANALYTICS IN FRAUD DETECTION AT PAYPAL

Arno Kolster
Sr. Database Architect
(Advanced Technology Group – Site Operations Infrastructure)

Image from Boris Müller's "Visual Poetry 6" (http://www.esono.com/boris/projects/poetry06/visualpoetry06/)

HPC USER FORUM

September 19, 2012

PayPal™

THE PROBLEM

Detecting fraud in 'real time' as millions of transactions are processed between disparate systems at volume.

Finding suspicious patterns that we don't even know exist in related data sets.

Ability to create and deploy new fraud models into event flows quickly and with minimal effort.

Provide environment for fraud modeling, analytics, visualization, M/R, dimensioning and further processing.

THE CHALLENGES

5 9s availability, scalability and reliability in a 24x7x365 environment. “PayPal is always open” *.

How to keep fraud models current and ensure integrity of incoming events and data.

Maintaining a graph of identities, transactions, bank accounts, credit cards, ips, etc. to support the models.

Keep Operations simple. Small team of SAs and DBAs.

Educate peers and higher ups of new technology and concepts so they ‘get it’. “HPC what?”

WHAT KIND OF VOLUME?

10 million+ PayPal logins / day.

13 million financial transactions / day.

300 variables calculated per event for some models.

~4 Billion inserts / day.

~8 Billion selects / day.

OUR SOLUTION - TRINITY

Real time linking platform for identities from various source systems. Built a social network.

Highly distributed open source databases for OLTP storage of edges and links. Architected for scale out and HA.

Intelligent gateways, message routing & delivery to heterogeneous systems.

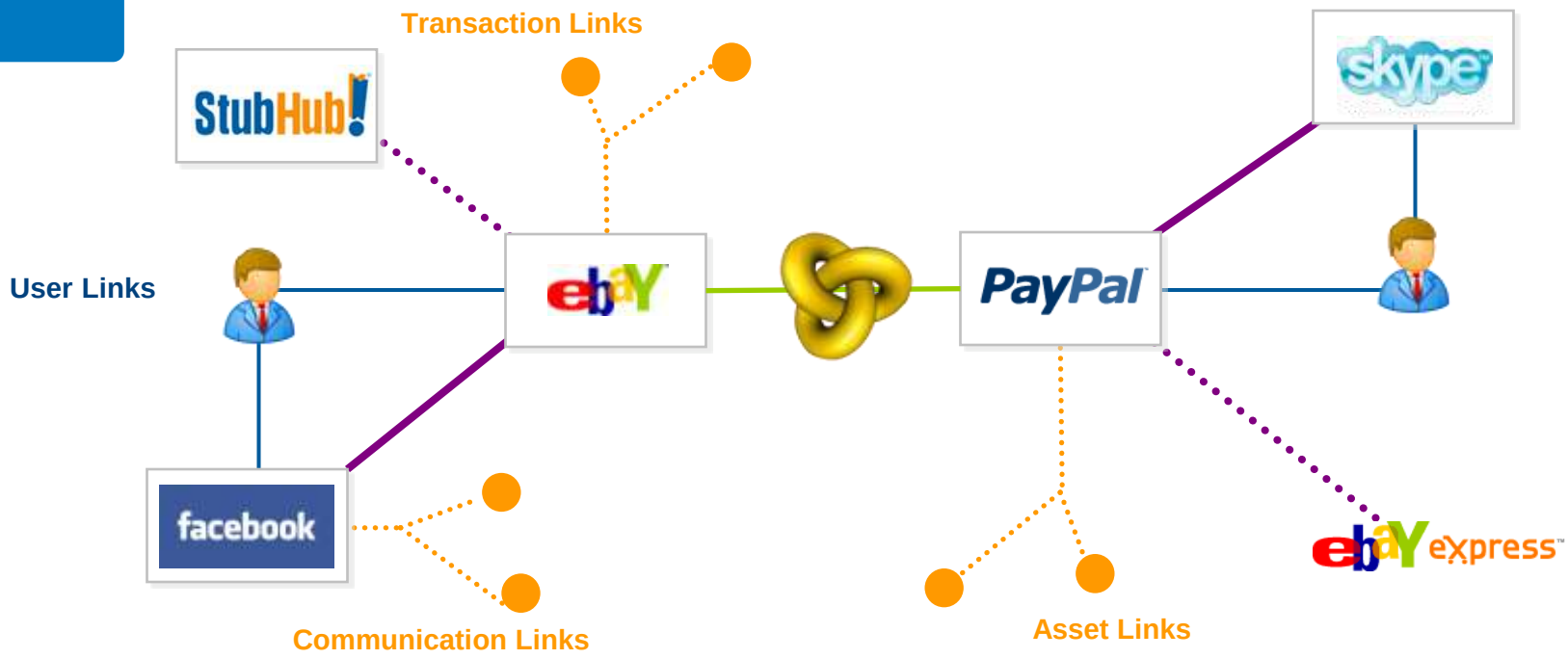
Inline stream analytics using CEP and ESP.

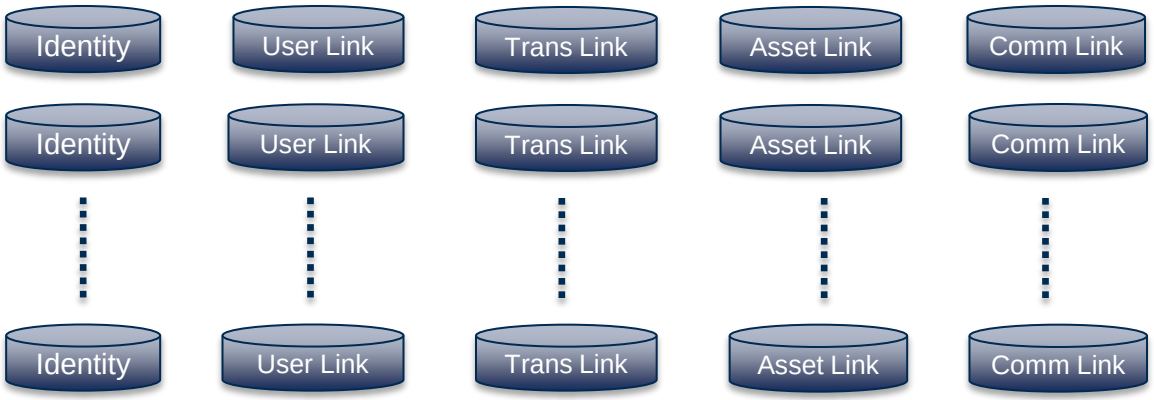
Downstream analytics environments for further processing.

Leveraged HPC architecture and hardware where it makes sense.

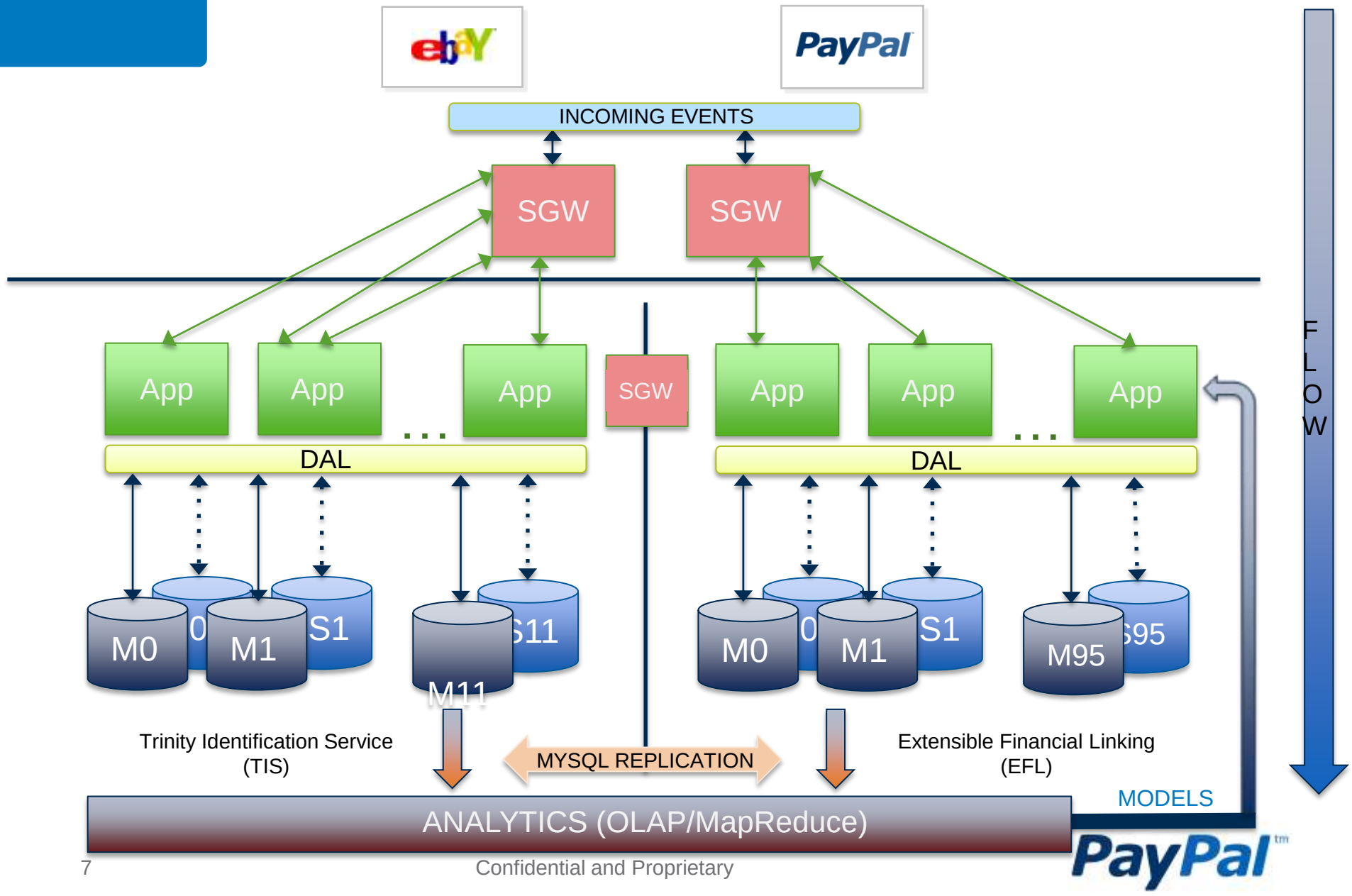
Standardized operations – h/w, s/w deployment, monitoring, command & control processes, etc.

TRINITY IDENTITY SERVICE (TIS) LINKING PLATFORM



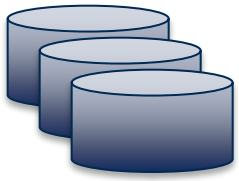


HIGH LEVEL SYSTEM OVERVIEW



HOW DID WE MANAGE TO SCALE ?

TRINITY DB PROLIFERATION



36 Instances

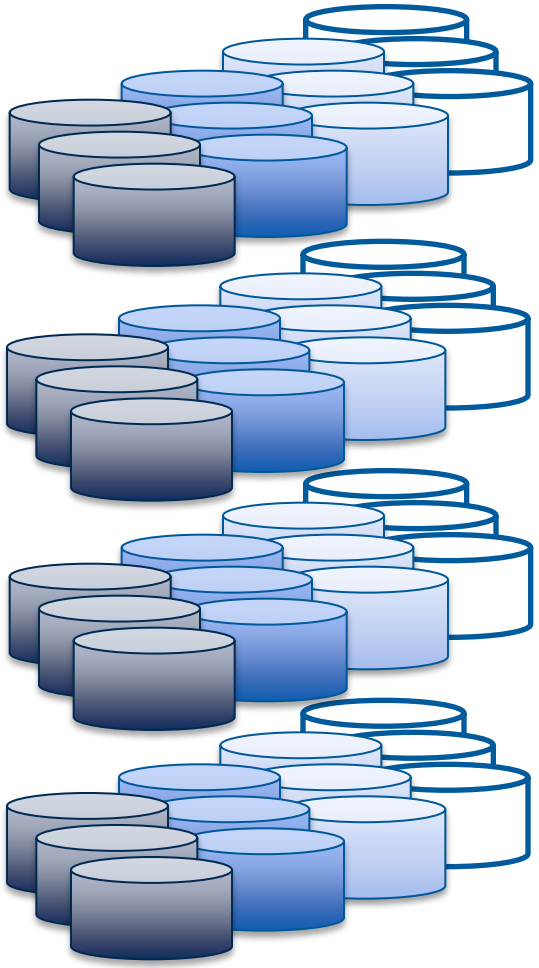
12 Masters, 24 ROs

TIS, TAS

12 shards

3 DBAs / 4 SAs

2007



1800 Instances

600 Masters, 1200 ROs

8 Billion Selects/Day

TIS, TAS
4 Billion Identities
Billions of Links

ARS, NEO, UVS, NA

EFL

Horizontal scaling

3 DBAs / 4 SAs

2012

PayPal

DERIVATION OF MODELS

Metrics / variables / summaries generated from inline processing of events using CEP (>300 metrics / event)

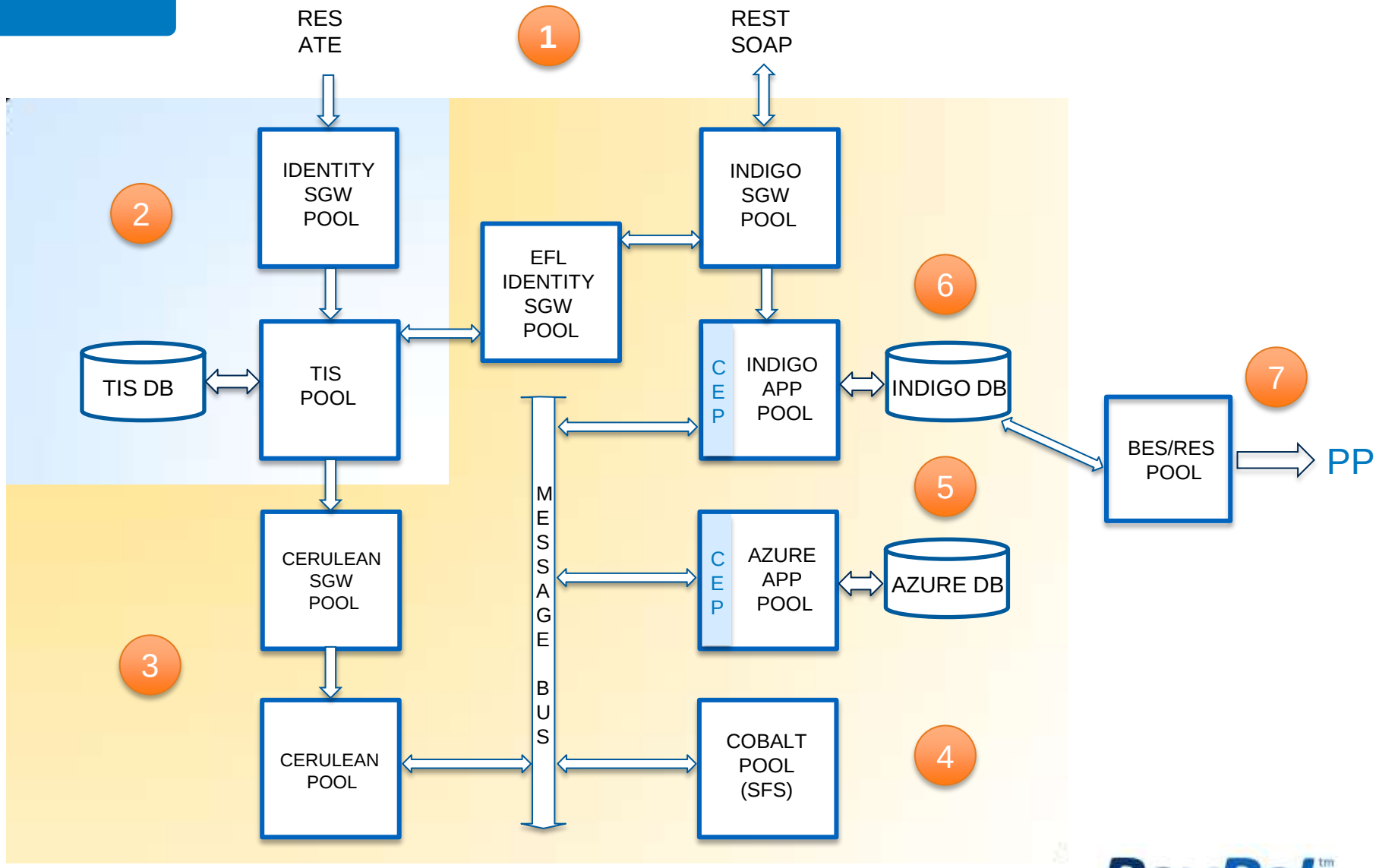
Scoring of events based on historical and current metrics.

Scores sent on to PayPal flows for further RISK modeling or transaction blocking.

Different Fraud Models generate different types of scores.

New Fraud Models created based on success ratio of previous ones or reaction to change in data and usage patterns. (R, SAS, M/R, vectoring)

EFL (EXTENSIBLE FINANCIAL LINKING)



PayPal

WHERE ARE WE USING HPC?

Infiniband on all internal Trinity network (Mellanox QDR 40Gb dual plane)

SGI InifiniteStorage IS4600 for EFL databases.

3 SGI Altix ICE 8200/8400 clusters for all 120+ EFL memory based apps – no disk i/o overhead.

MPI “like” apps. MPP features with scale out and affinity processing.

R&D with Lustre on Hadoop cluster and POC of columnar based database.

SGI ALTIX ICE 8200/8400

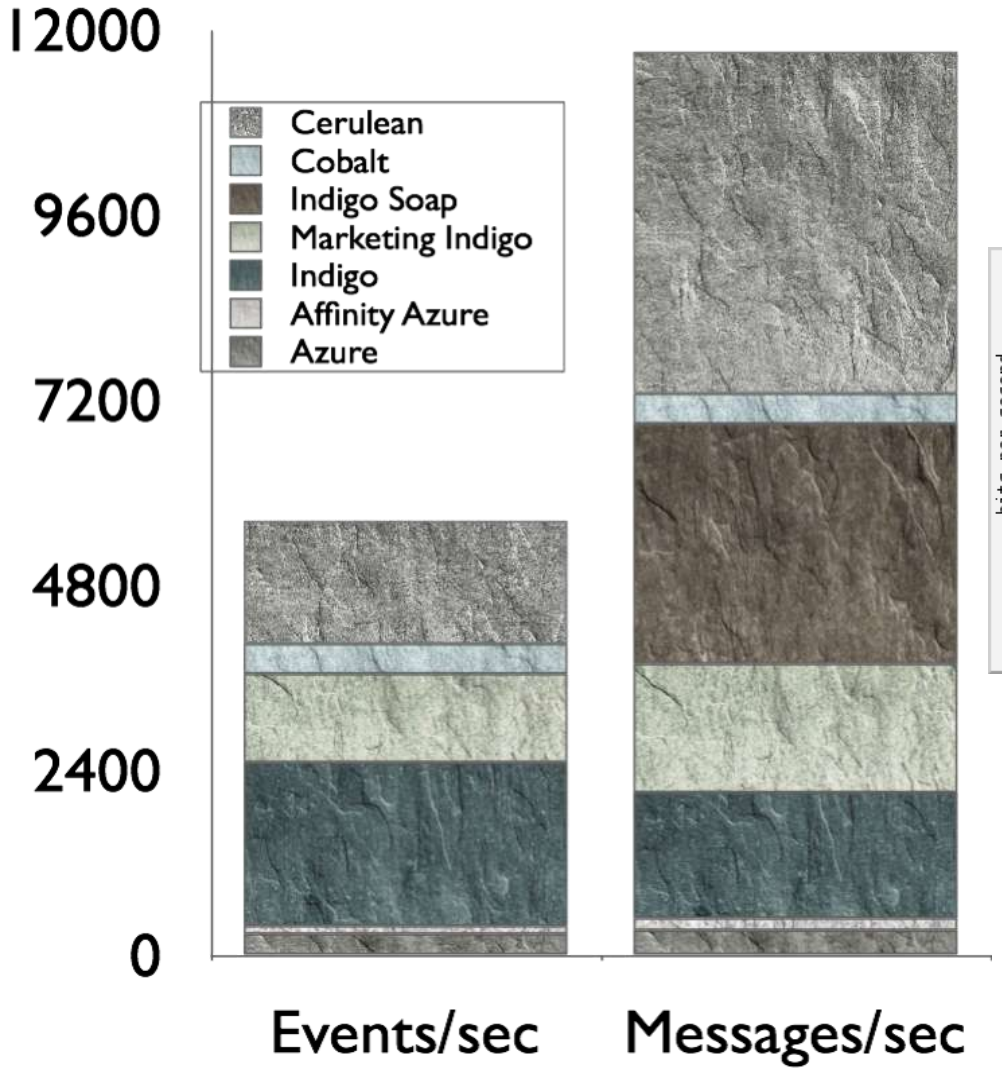
Supports multiple deployment strategies in the same cluster

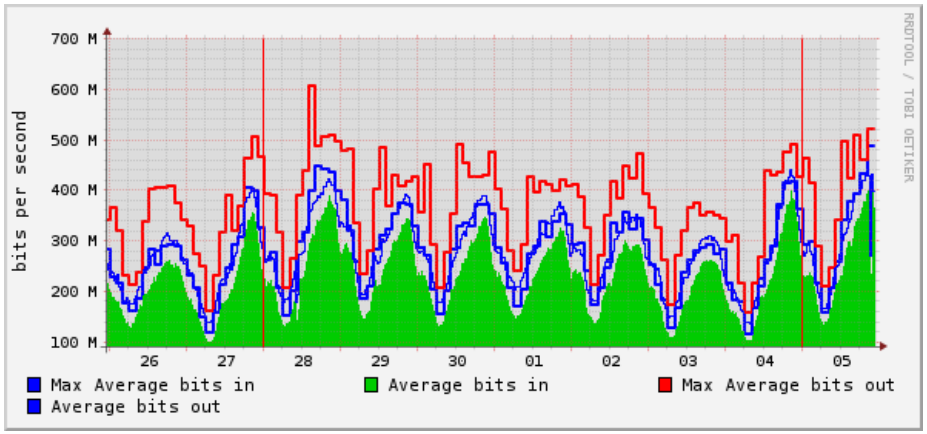
SFS	SFS	SFS	SFS	SCORING	SCORING	SCORING	SCORING
SCORING	SCORING	SCORING	SCORING	SCORING	SCORING	SCORING	SCORING
SCORING	SCORING	MSGING	MSGING	MSGING	MSGING	MKTG	MKTG
MKTG	MKTG	MKTG	MKTG	REPLAY	REPLAY	REPLAY	AFFINITY
AFFINITY	AFFINITY	AFFINITY	AFFINITY	RDBMS	RDBMS	RDBMS	RDBMS
RDBMS	RDBMS	RDBMS	RDBMS	RDBMS	RDBMS	RDBMS	RDBMS
RDBMS	RDBMS	RDBMS	RDBMS	ANALYSIS	ANALYSIS	METRICS	METRICS
METRICS	METRICS	METRICS	METRICS	API	API	API	API

EFL Cluster Provisioning By Application

PayPal™

EFL MESSAGING STATISICS
CYBERWEEK 2011





THE FUTURE

Add a true graph database to complement link databases.

Increase performance SLAs through h/w upgrades (SSDs?).

Add true analytics database that is updated in real time.

Change key/value type datastore from structured to semi-structured.

Keep educating peers and higher ups so they ‘get it’. “HPC? Yes!”

Sell one like this

NVIDIA TESLA S870 x3 3U 18GB 12x C870 HPC CUDA GP-GPU COMPUTE CLUS

Like | Want | Own

Item condition: **Used**

Price: **US $999.99**

Buy It Now

Add to cart

Best Offer:

Make Offer

Add to Watch list

BillMeLater **Spend $899+ & get 18 months financing**
Subject to credit approval. See terms

Shipping: **$100.00** Economy Shipping | See details
Item location: **Spring, Texas, United States**
Ships to: **United States**

Delivery: Estimated between **Thu. Sep. 20 and Wed. Oct. 3**

Payments: PayPal, Bill Me Later | See details

Returns: No returns or exchanges, but item is covered by eBay Buyer Protection.

eBay Buyer Protection
Covers your purchase price plus original shipping.
Learn more

AKOLSTER@PAYPAL.COM