




Scaling Up: Using JBoss Clustering  
for Large-Scale Information Delivery

James Dixon, CTO, Pentaho Corp  
June 2006





### Contents

- Introduction
- Reasons for large-scale information delivery
- Different approaches
- Scalability Factors
- Clustered Architecture
- Cluster Performance
- Summary

### Introduction

This presentation shows how to create an incrementally scalable architecture for large-scale information delivery using JBoss AS, JBoss JGroups, Apache, and open source reporting components from Pentaho.






### Information Delivery Evolution

**Traditional Business Intelligence (BI)**



- Stage 1 – Paper-based chaos
- Stage 2 – Microsoft-based chaos
- Stage 3 – Standard Reporting
- Stage 4 – Advanced Reporting
- Stage 5 – Dashboards and Analysis
- Stage 6 – Datamining

Operational BI

### Standard Reporting



- Bursting
- Exception Reporting
  - Reduces information overload and resource utilization
- High internal reach
- Supports best practices

### Standard Reporting

**Bursting**

- Fan-Fold finger-flicking history
- Modern incarnation
  - Same process, fewer interns and injuries
- 'Walrus' Approach
  - Single large process
- 'Swarm' Approach
  - Multiple small processes

## Standard Reporting

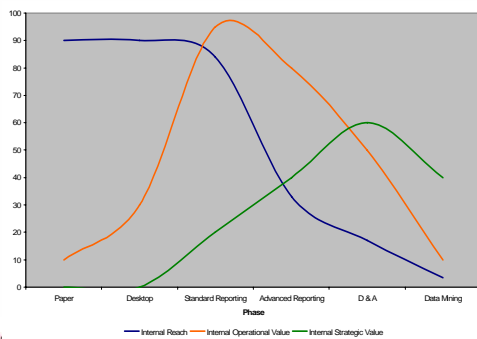
### Operational Business Intelligence

- Process-centric
- Transactional triggers
- High internal reach
- Inline data integration / Enterprise Application Integration

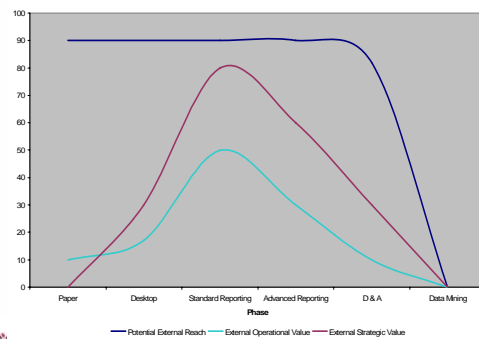
## Advanced Reporting

- Ad-hoc/Parameterized Reporting
- Customized Subscriptions

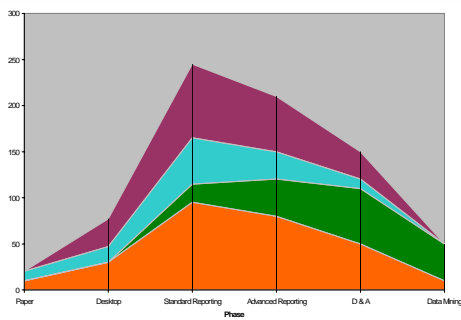
## Internal Information Delivery



## External Information Delivery



## Value of Information Delivery



## Large-Scale Information Delivery

### Standard Reporting

- Bursting
- Operational BI

### Advanced Reporting

- Scheduled delivery of subscribed reports

### Scale Up: 'Walrus' Approach

Single large process

Pros:

- Potentially less database access

Cons:

- Hard to control
- Hard to scale incrementally
- Hard to distribute over time or machines



### Scale Out: 'Swarm' Approach

Many iterations of a simple process

Pros:

- Easy to control
- Easy to scale incrementally
- Easy to distribute
- Works equally for Bursting, Operational BI and Subscriptions

Cons:

- Potentially more database access

Bursting/Operational BI (BOBI) Architecture



### Scalability

Factors that affect scalability

- Data source
  - Transactional databases, data warehouse
- Content generation
  - PDF / Excel / HTML generation
- Content delivery
  - Email servers, network bandwidth etc.
- Network connections between servers

Any of these factors can be limiting



### Scalability: Data Sources

- Transactional data-sources typically not indexed for report generation
- Data marts can be segmented



### Scalability: Delivery

**Email server**

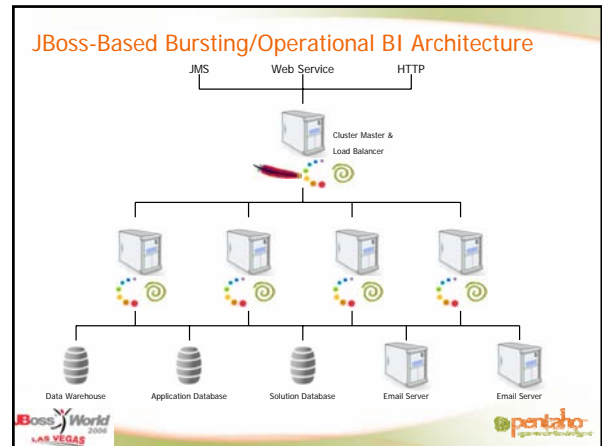
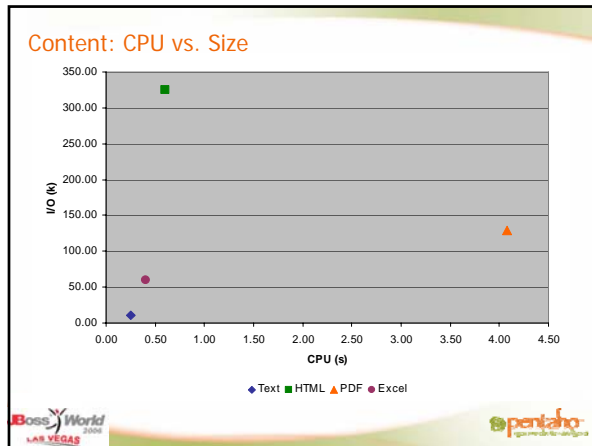
- Often a bottleneck
- Bottleneck is I/O: content size matters
- Can be clustered
- Bypass spam and anti-virus components
- Might need to throttle content generation



### Scalability: Content Generation

- Computationally expensive
- Different costs for PDF / HTML / XLS / TXT

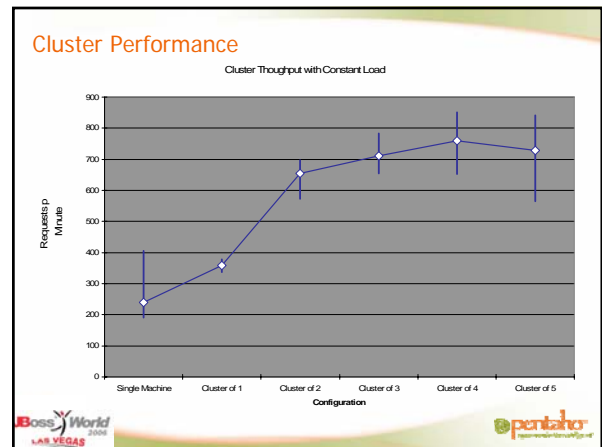


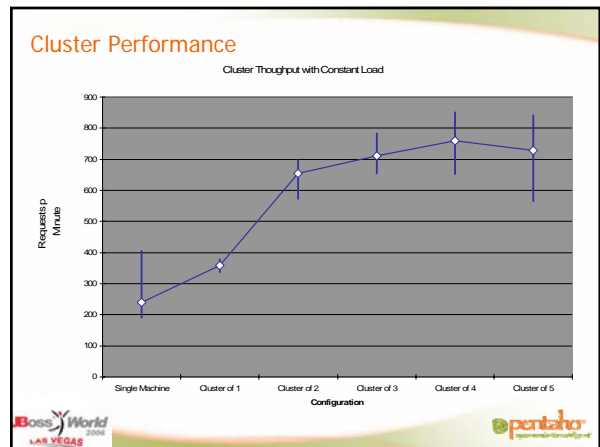
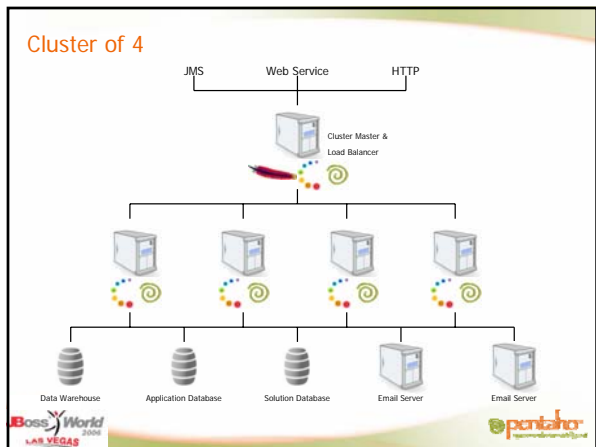
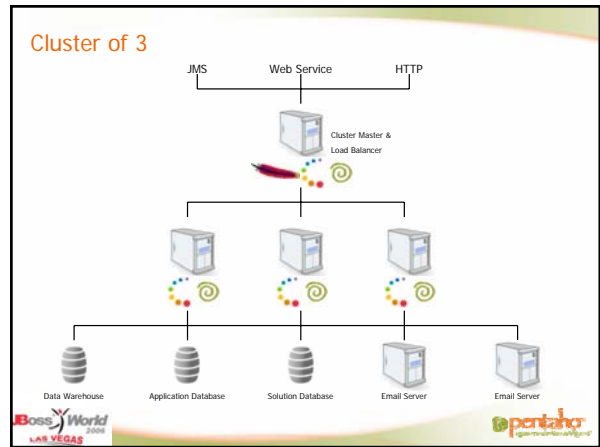
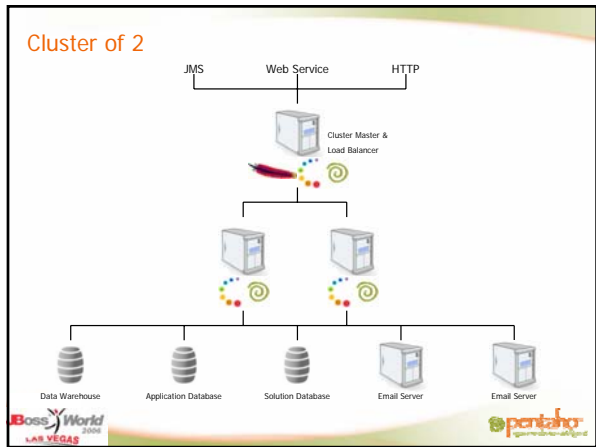
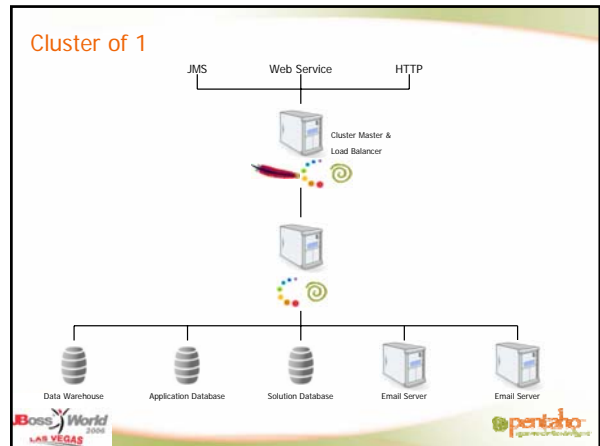
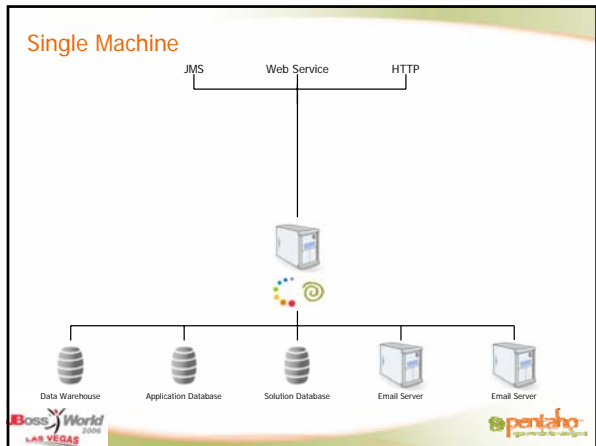


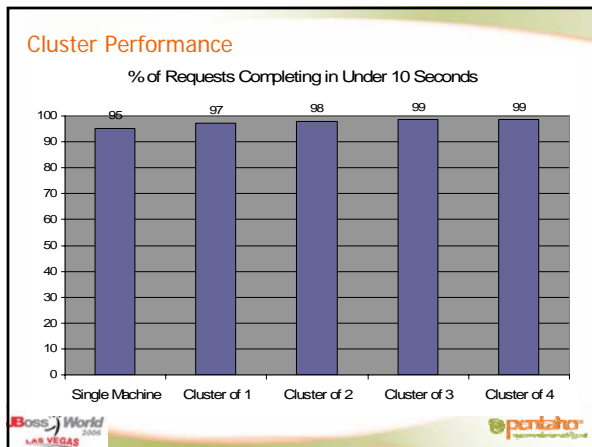
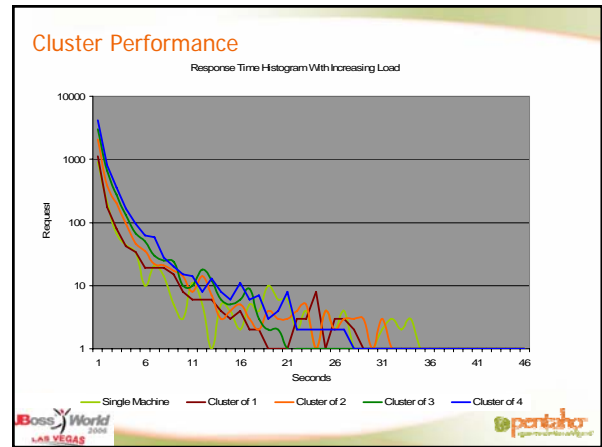
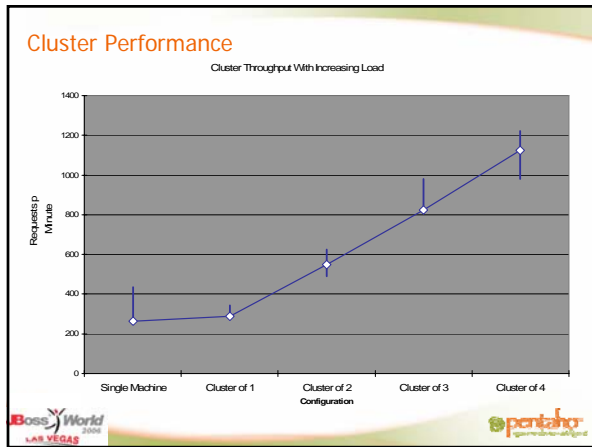
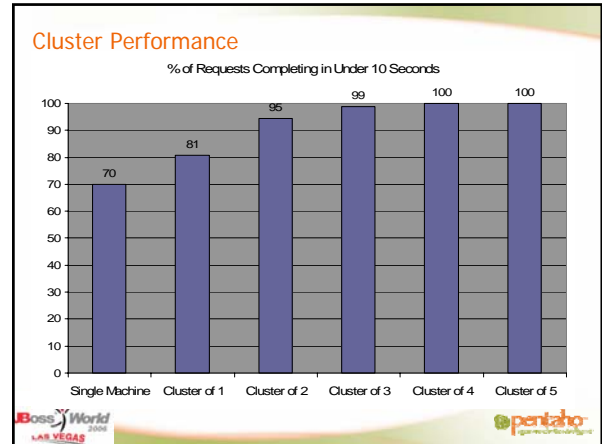
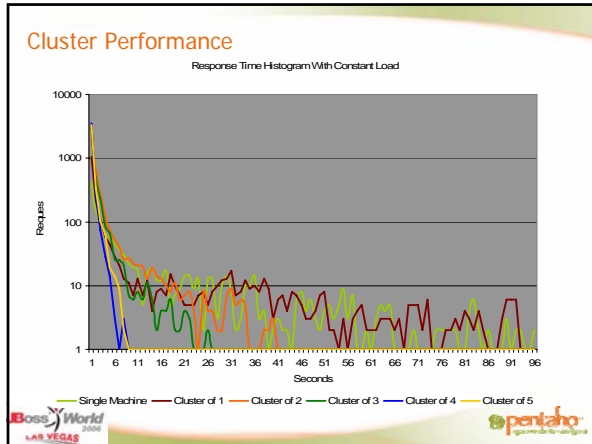
- ### JBoss-Based Bursting/Operational BI Architecture
- BOBI Tier
    - JGroups cluster
    - JBoss AS
    - Pentaho Report Server
  - Cluster master
    - Apache HTTP server 2.0.58 with mod\_jk module version 1.2.15
    - JGroups cluster master
  - JMS / Web Services for Operational BI

- ### JBoss-Based Bursting/Operational BI Architecture
- Cluster configuration details:
- Single CPU
  - 2 GB RAM
  - JBoss AS 4.0.3
  - JBoss JGroups?
  - Pentaho 1.1.5

- ### Cluster Performance
- Load / throughput testing
- With constant load as cluster size increases
  - With increasing load as cluster size increases
  - HTML, PDF, Excel content generation
  - Relational, XML, OLAP data sources
  - 'Measure what you want'
  - TPC-W used to simulate concurrent requests and measure client's response times







### Summary

JBoss JGroups and Pentaho provide linear scalability for content generation for large-scale information delivery using a Bursting / Operation Business Intelligence architecture.

JBoss World 2008 LAS VEGAS pentaho

## Resources and Links

### Pentaho

Software: <http://www.pentaho.org/downloads>  
(whitepapers, binaries, source code etc)

Biz Dev: [dhenry@pentaho.org](mailto:dhenry@pentaho.org)

CTO: [jdixon@pentaho.org](mailto:jdixon@pentaho.org)



## Resources and Links


### JBoss

JGroups:

[http://labs.jboss.com/portal/index.html?ctrl:id=page\\_default.info&project=jgroups](http://labs.jboss.com/portal/index.html?ctrl:id=page_default.info&project=jgroups)

JBoss AS:

[http://labs.jboss.com/portal/index.html?ctrl:id=page\\_default.info&project=jbossas](http://labs.jboss.com/portal/index.html?ctrl:id=page_default.info&project=jbossas)



Scaling Up: Using JBoss Clustering  
for Large-Scale Information Delivery

James Dixon, CTO, Pentaho Corp  
June 2006

