Cray XC30 Hadoop Platform

Jonathan (Bill) Sparks
Howard Pritchard
Martha Dumler

Safe Harbor Statement

This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

Hadoop MapReduce: new workload for XC30

- **Apache Hadoop is an open-source implementation of a "big data" application framework pioneered by Google.**
  - Typical commercial applications are things like
    - Social networking, "web 2.0" applications.
    - Business Intelligence – targeted marketing, sales analysis
    - Inventory and sales pattern analysis
- **HPC applications in areas like genomics and seismic (search) and generally in post-processing large data sets.**

Hadoop is a new/additional workload for your XC30

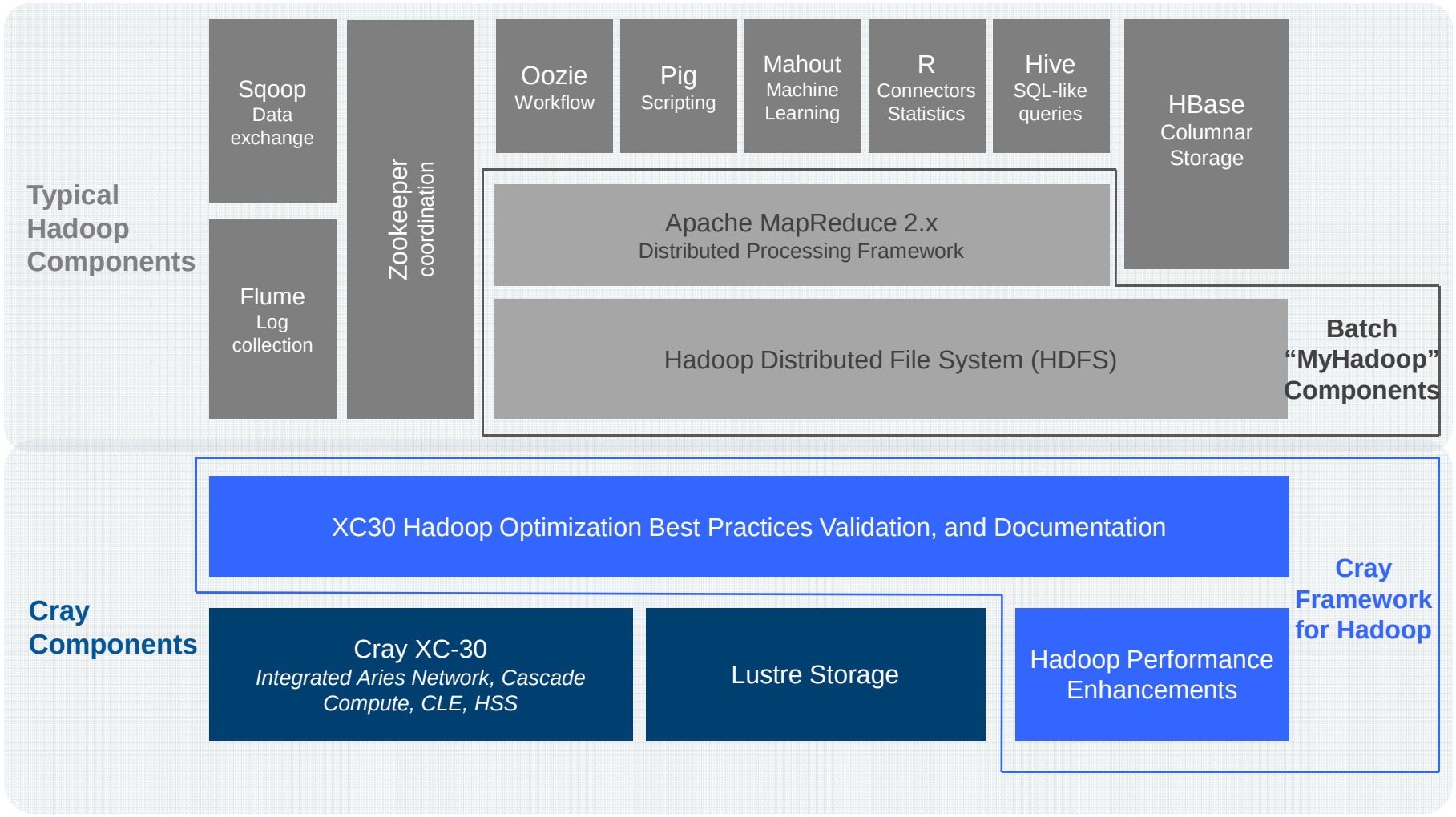
XC30 Hadoop Releases

- **Aimed at existing XC30 customers that want to run hadoop on a subset of their compute nodes**
- **Provided to assist and optimize with installation and getting familiar with the hadoop environment and develop applications that take advantage of hadoop**
- **Download and go**
  - No cost
  - No formal support contract

HPCS Challenges

- **Workload Management**
  - Some Hadoop components are long-running, shared services
    - HPCS is generally a batch-oriented platform
  - Hadoop components that are batch-oriented (map-reduce/YARN) don't integrate well with traditional HPCS WLMs
- **Storage and I/O**
  - Hadoop traditionally uses local, persistent storage
  - Hadoop Distributed File System (HDFS) tightly coupled with many hadoop components
    - HPC Systems utilize global parallel filesystems

Cray XC30: Cray Framework for Hadoop

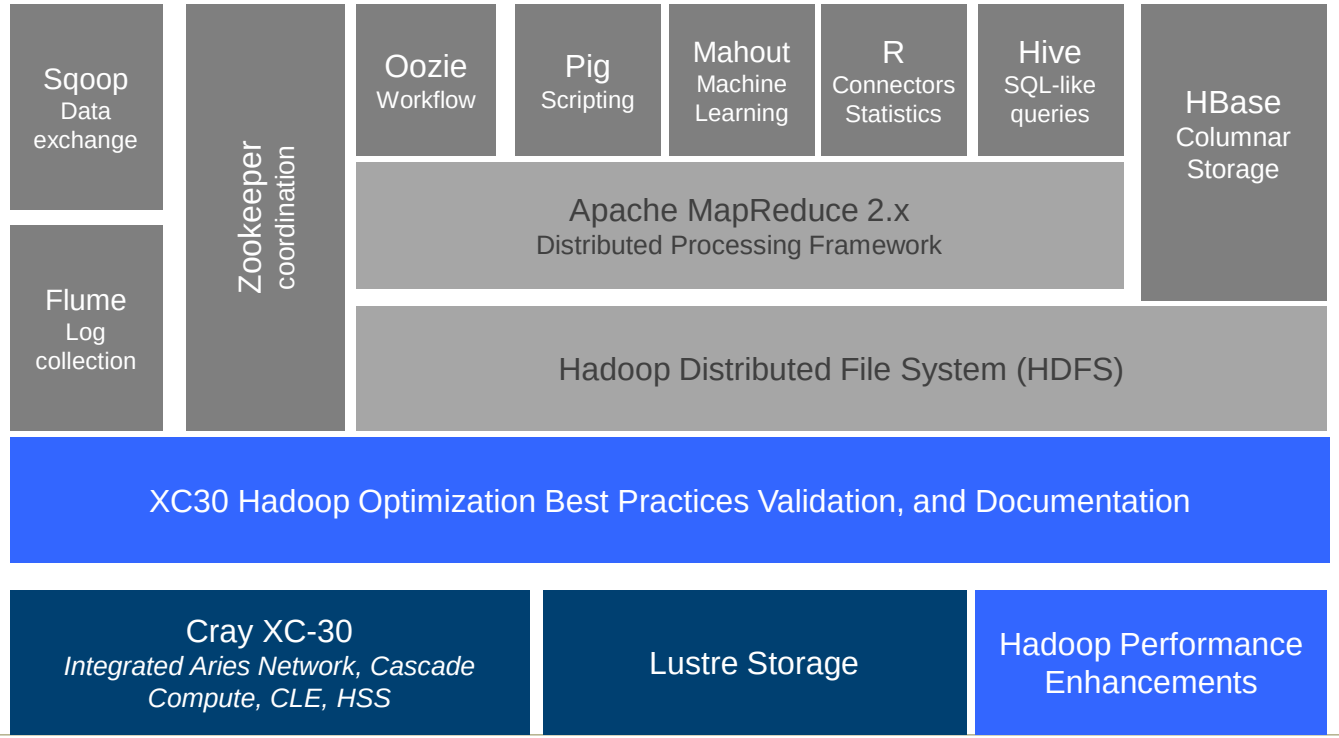


XC30 Hadoop Initial Release

December 2013

XC30 Hadoop - Initial Release

- **System Components:**
  - Batch template files for Moab/Torque, PBSPro
  - XC30-specific configuration settings
  - Batch Components: MapReduce, YARN, HDFS
  - Plus “environment” components (Hive, Flume, HBase, Zookeeper, …)



COMPUTE | STORE | ANALYZE

Initial Release Contents

- **Site chooses Hadoop distribution – Cray provides XC/XE instructions**
  - Apache, Bigtop, Cloudera, HortonWorks, IDH – all used internally
  - Hadoop "environment" packages optionally installed on re-purposed compute nodes
- **Suggested Site Customizations**
  - Adjust YARN parameters to reflect number of cpus/node and mem/node
  - Adjust default memory limits for mapper and reducer tasks

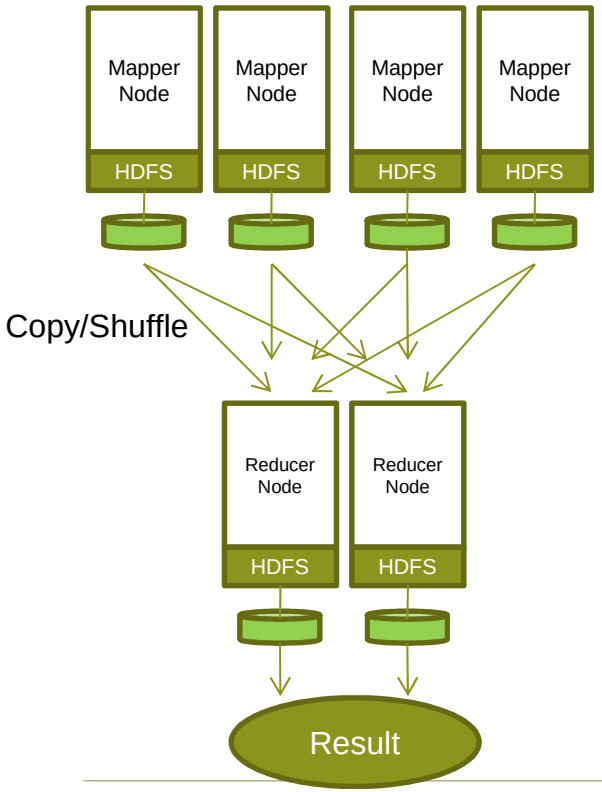
Best when all nodes configured for same cpus/node and mem/node

- **Batch script and template files provided for quickstart examples**
  - PBS Pro
  - MOAB/Torque

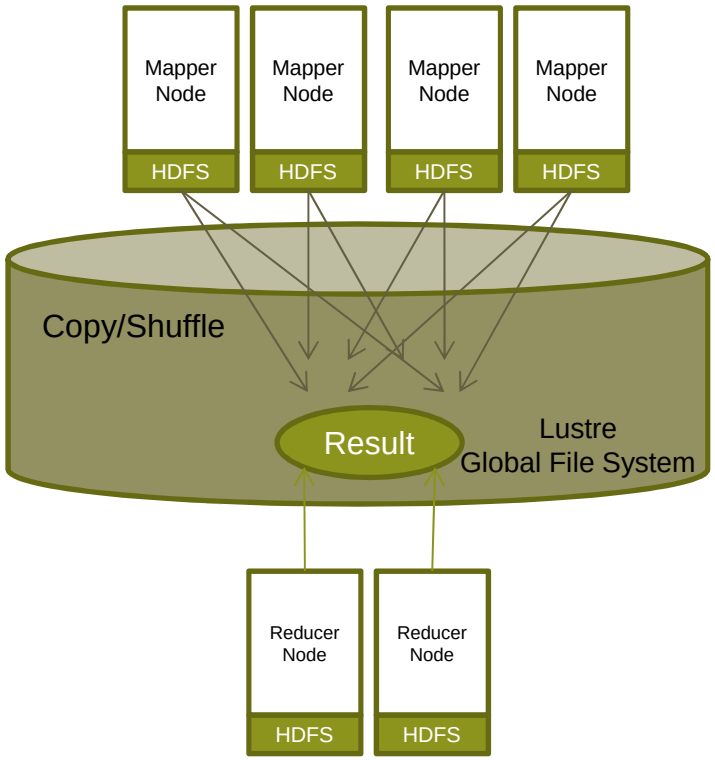
MapReduce on XC30

- Utilizes Global Filesystem – single copy of input data
- XC30 High-speed Aries Network – fast data movement
- Large compute farm – easy to install/configure additional MR nodes

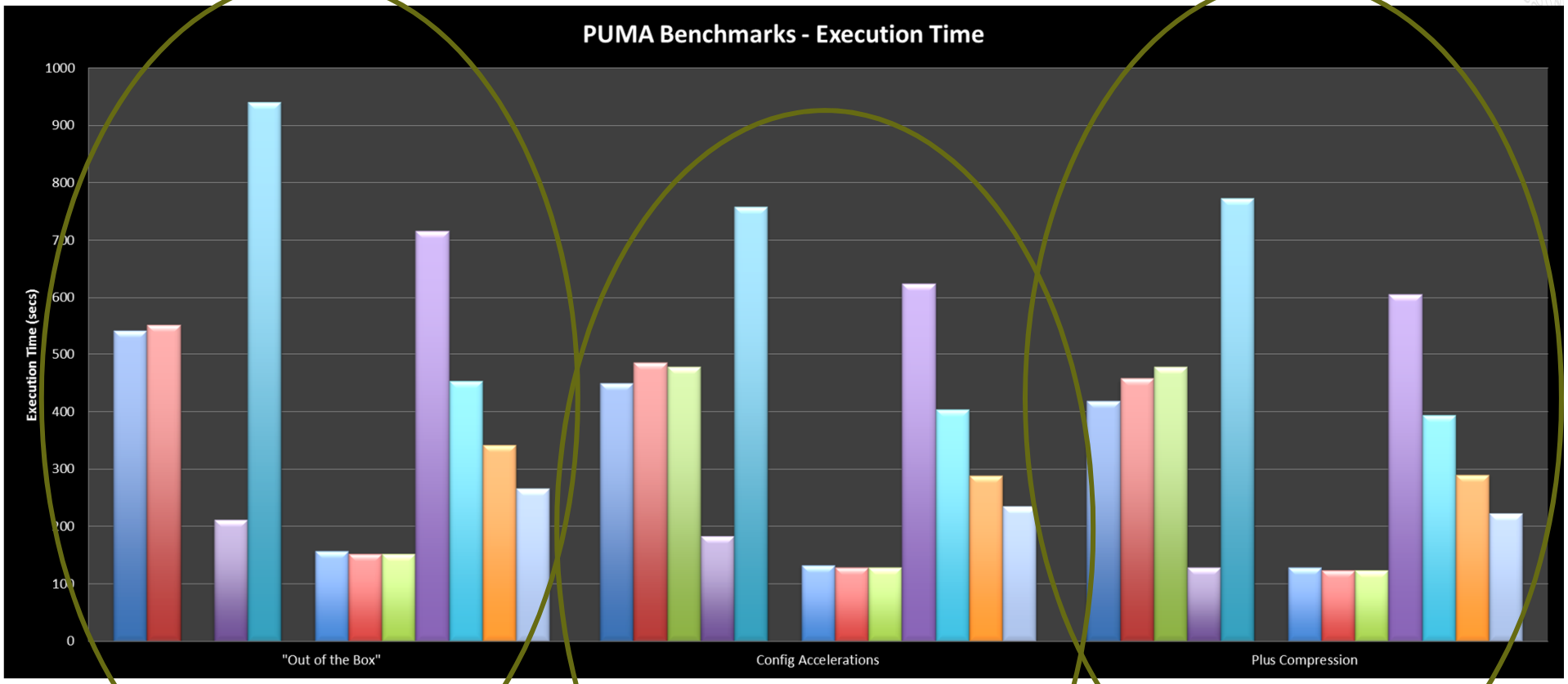
Typical MapReduce



XC30 MapReduce



PUMA Execution Details



Performance Improvement – XC30 Tuning

- Achieved an average of **20% performance improvement** of the PUMA benchmarks with XC30-specific tuning as compared to hadoop "out of the box"

XC30 Hadoop Release Update

April 2014

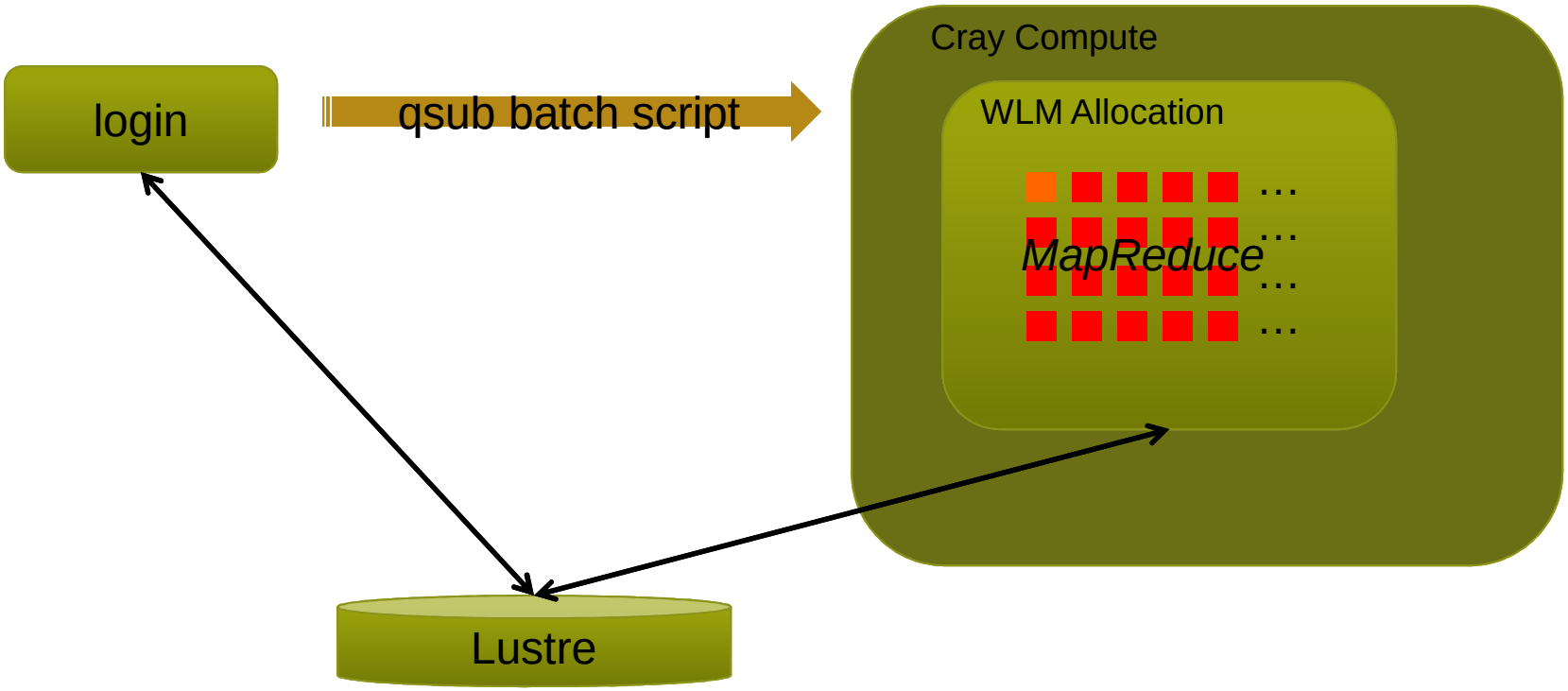
Update Release – April 2014

- **Focus on batch "myhadoop"**
  - Tested at scale on XC30 and XE
  - Support esLogin/JDK1.7 compatible
  - Update to myHadoop (Apache 2.0.6)
    - Simplified batch interaction
      - Batch script reduced from 130 lines to 34 lines of code
    - Dynamic parameter selection based on job size
    - Hadoop rawfs support for native Lustre/Filesystem
    - Reducer Shuffle update
    - Additional Hadoop instructions (pig, mahout, crunch)
  - Updated Documentation
- **Avoiding Lustre problems….**

Cray myHadoop

- **Uses system batch system to create on-demand Hadoop instance.**
- **Works with PBS, Moab/Torque**
- **Includes Hadoop, Pig (scripting), Mahout (Machine Learning), Crunch (pipelining)**
- **Easy setup and install**
  - Pre-configured
  - Dynamic configurations matching node count
  - HDFS or noHDFS configuration switch

myHadoop



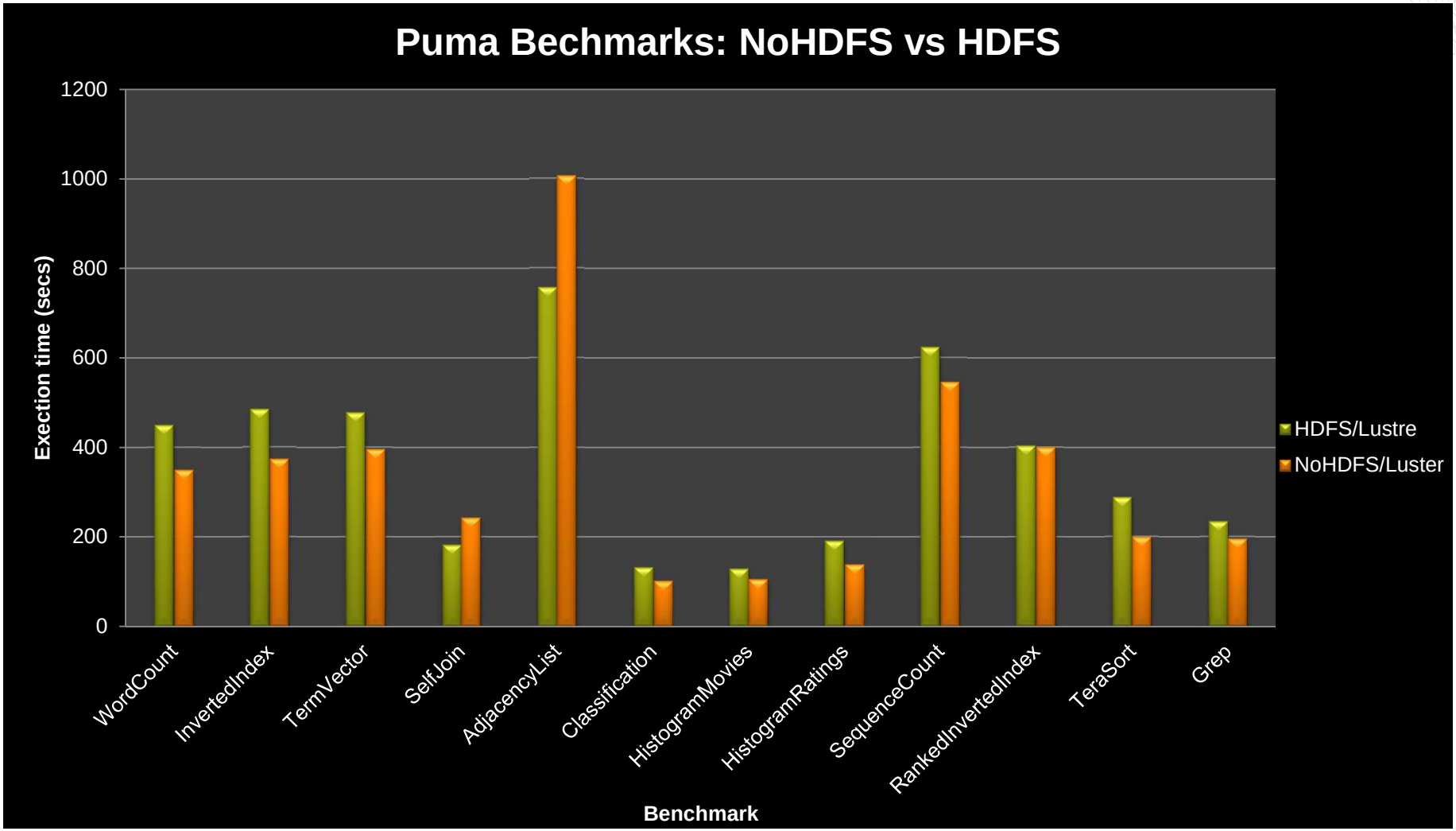
Hadoop ResourceManager – YARN Allocator
Hadoop NodeManager – YARN Task executor

HDFS and Lustre

- A typical Hadoop deployment, makes extensive use of HDFS:
  - Application jar files
  - job logfiles to HDFS
  - Input and output data, ….
- HDFS on top of Lustre
  - performance penalties
  - Larger jobs cause problems with Lustre software stack

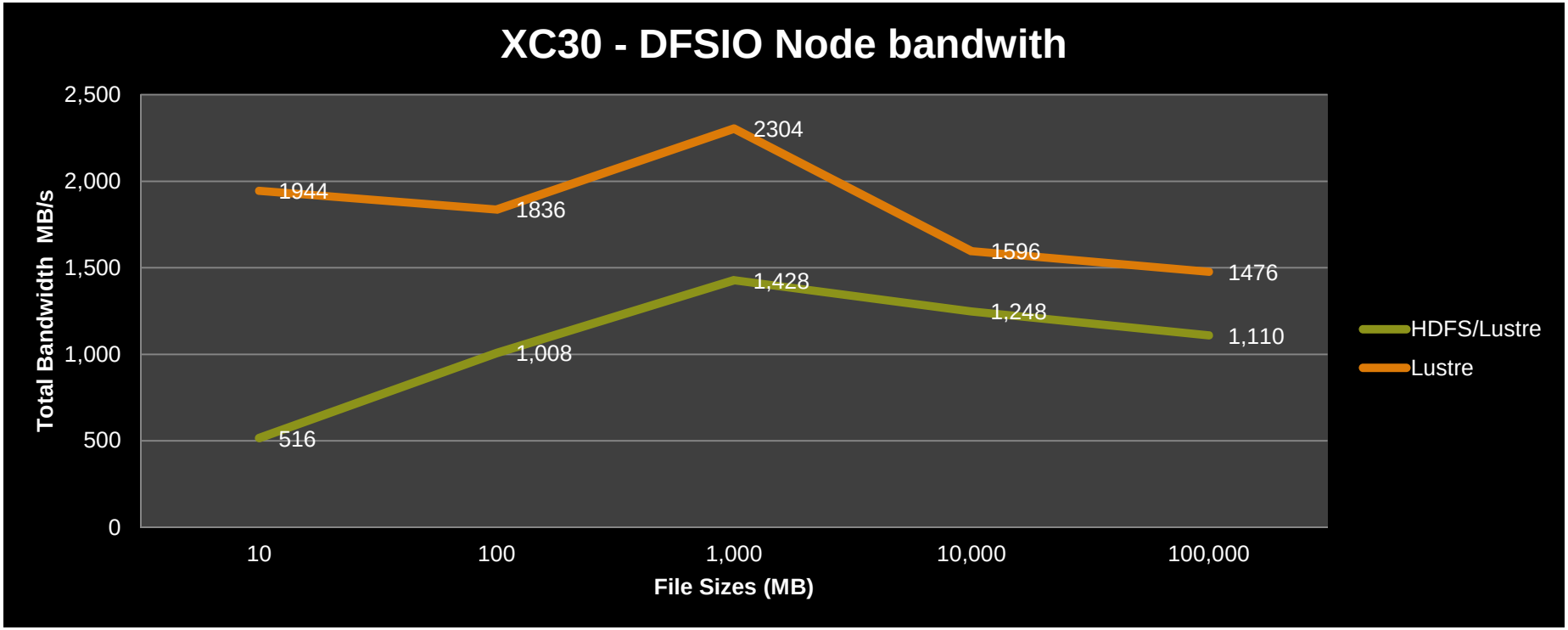
- **Solution:**
  - Default config parameters set to run without HDFS
  - Off-loading of files to tmpfs rather than writing to lustre FS

Puma Benchmarks HDFS/NoHDFS



Hadoop DFSIO test times

- **Single node – writing to HDFS/Lustre vs Lustre**
  - Throughput rate with/without HDFS



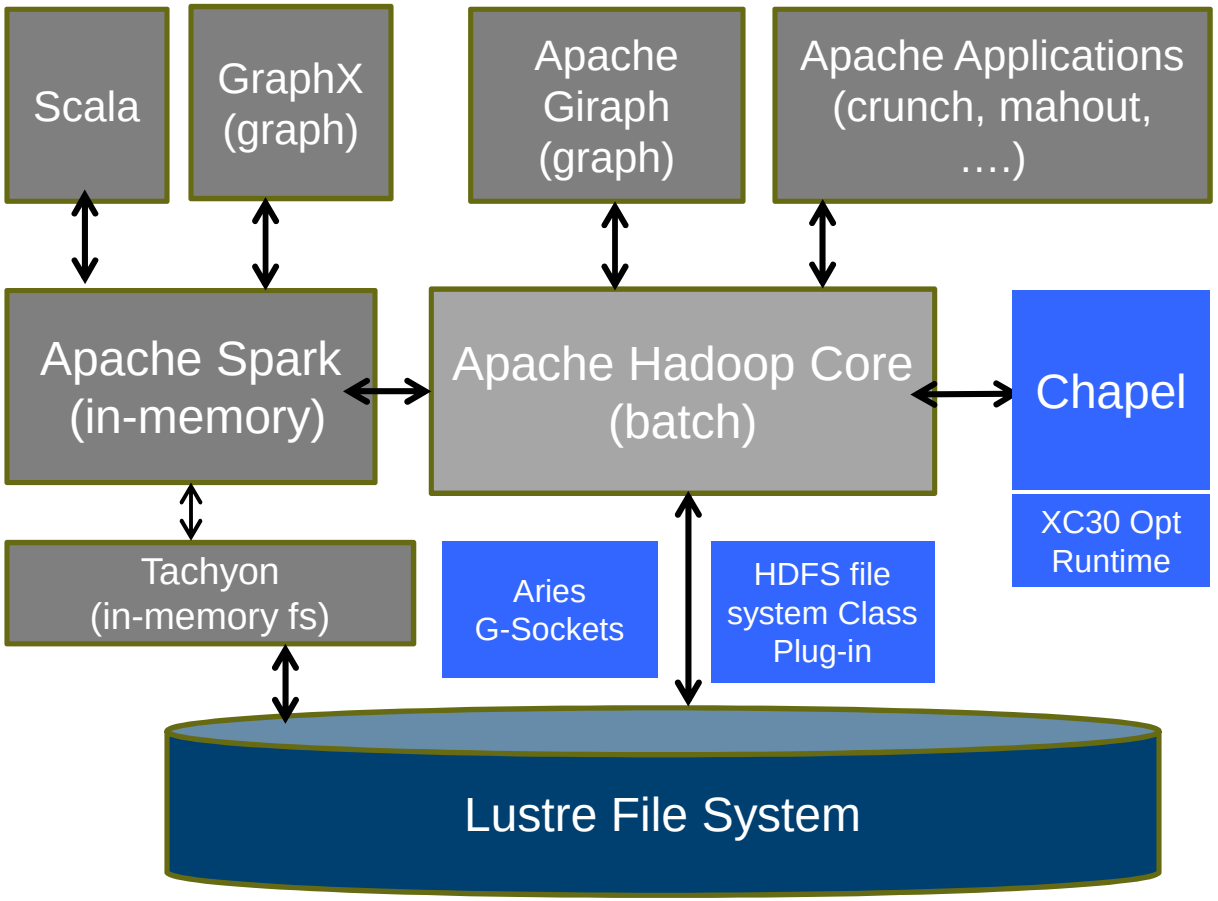
XC30 Hadoop Next Release

Date is TBD

Prototypes

- **Aries RDMA -** protocol over RDMA that supports a socket-level API for applications
  - Transparent bandwidth advantage without modifying a TCP sockets based application
  - Bandwidth comparison:
    - TCP sockets: ~25Gbps
    - Gsockets: ~60 Gbps
- **Lustre Filesystem class – "native" Lustre Access**
- **Use of intermediate storage**
- **Integration with System Management** – allows provisioning/finer-grain management of compute nodes to grow/shrink map/reduce nodes

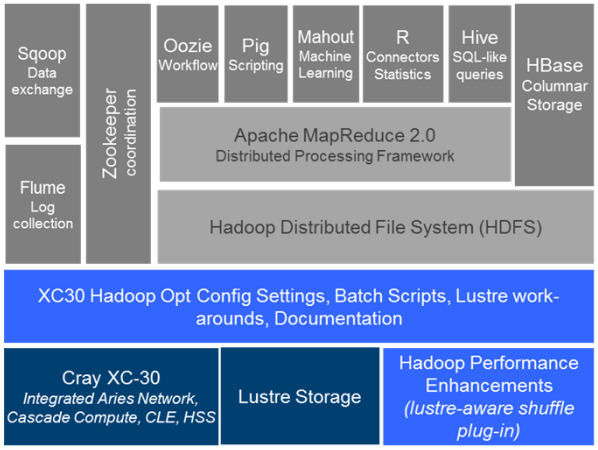
Additional Capabilities



- **Further capabilities**
  - Graph analytics
  - In-memory map reduce
  - Chapel with XC30 runtime
  - Tachyon (in-memory filesystem)

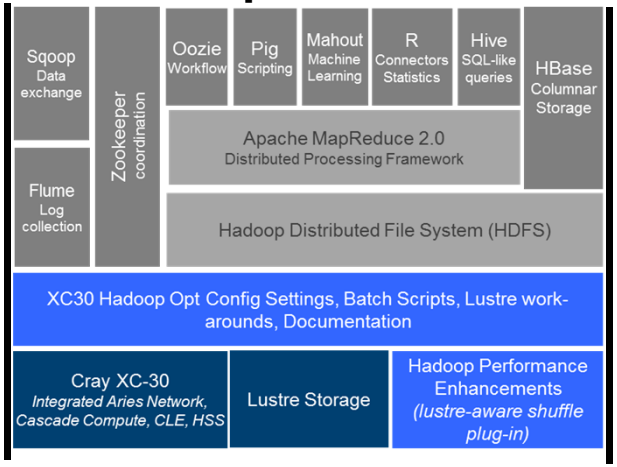
XC30 Hadoop Roadmap

LA Release Dec 2013



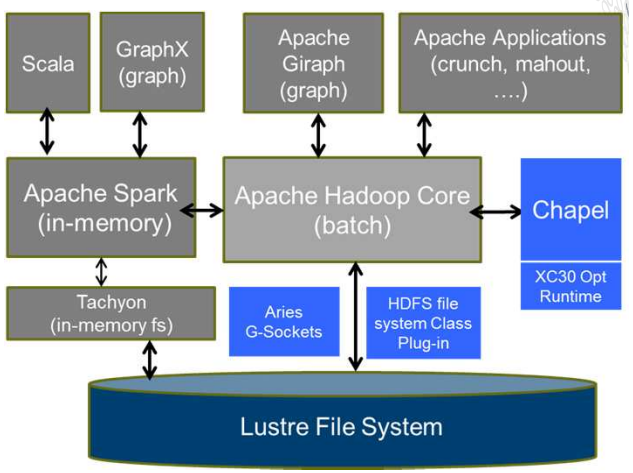
- XC30 config optimizations for 20% performance improvement
- Lustre-aware shuffle plug-in
- Batch “myhadoop”
- “native” hadoop for full env capabilities

LA Update April 2014



- Lustre fixes and workarounds
- Easier install process
- Focus on batch “myhadoop”

Proposed



- Lustre filesystem class plug-in (lustre “shim”)
- Additional capabilities:
  - In-memory map-reduce (spark)
  - Aries RDMA
  - GraphX
- Chapel HDFS I/O

Feedback

- What applications are you running or considering running with hadoop?
- What XC30 Hadoop roadmap components are being utilized?
- XE/XC Beta Sites providing feedback and sharing use cases via hadoop-dev@cray.com

Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, URIKA, and YARCDATA. The following are trademarks of Cray Inc.: ACE, APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.