Toward a Windows® Native Client (WNC)

Meghan McClelland
Meghan_McClelland@xyratex.com
LAD2013

xyratex

Overview

- At LUG 2013 there was expressed strong interest in a WNC client.
- Xyratex acquired IP from Oracle. The Lustre® code has changed since last updates, but material is still useful starting point.
- Funding proposal for evaluation being submitted for consideration by OpenSFS

Items in RED text in this presentation are asking for feedback

What is done today

- CIFS or pCIFS gateway
- NFS gateway
- Windows run in Linux hypervisor
- Maybe ok for staging small data

problems:

- limited numbers of gateways
  - (gateway bottleneck)
- coherency between gateways
- data flow through gateways
- additional layers of software

Current State

There is an existing prototype!

Implemented features and functions

- Mounting and unmounting
- Basic metadata operations
  - ls, create, rename, open, delete, attributes
- File I/O
  - direct (cached, mmap close but need work)
- LNET
- Byte range lock
  - flock
- Directory change notification (client only)
  - sorta like fcntl(F_NOTIFY) but with more details

Known Issues

- mmap lock implementation issues
- emulated page layer
- compiler issues

- testing - need ‘common’ Windows software

TODO?

- Branch code sync
- read-ahead
- CIFS re-sharing
- xattr & security support (ACLs?)
- Links / symlinks support
- Params-tree port
- lctl and lfs porting
- documentation
- Various codepage support (UTF8, gb2321)
- testing - partial acc-small port (at least sanity)? Or common Window software?
- GUI Tools
- mmap conflict detection
- cached I/O

Risks

- Caching
- mmap hooking
- Lustre filesystem constantly changing
- Code landing

- Licensing

Solution Alternatives

- Run Windows in hypervisor on Linux
- Export via SMB
- Export vis NFS
- pCIFS directly to OSS

Feedback!

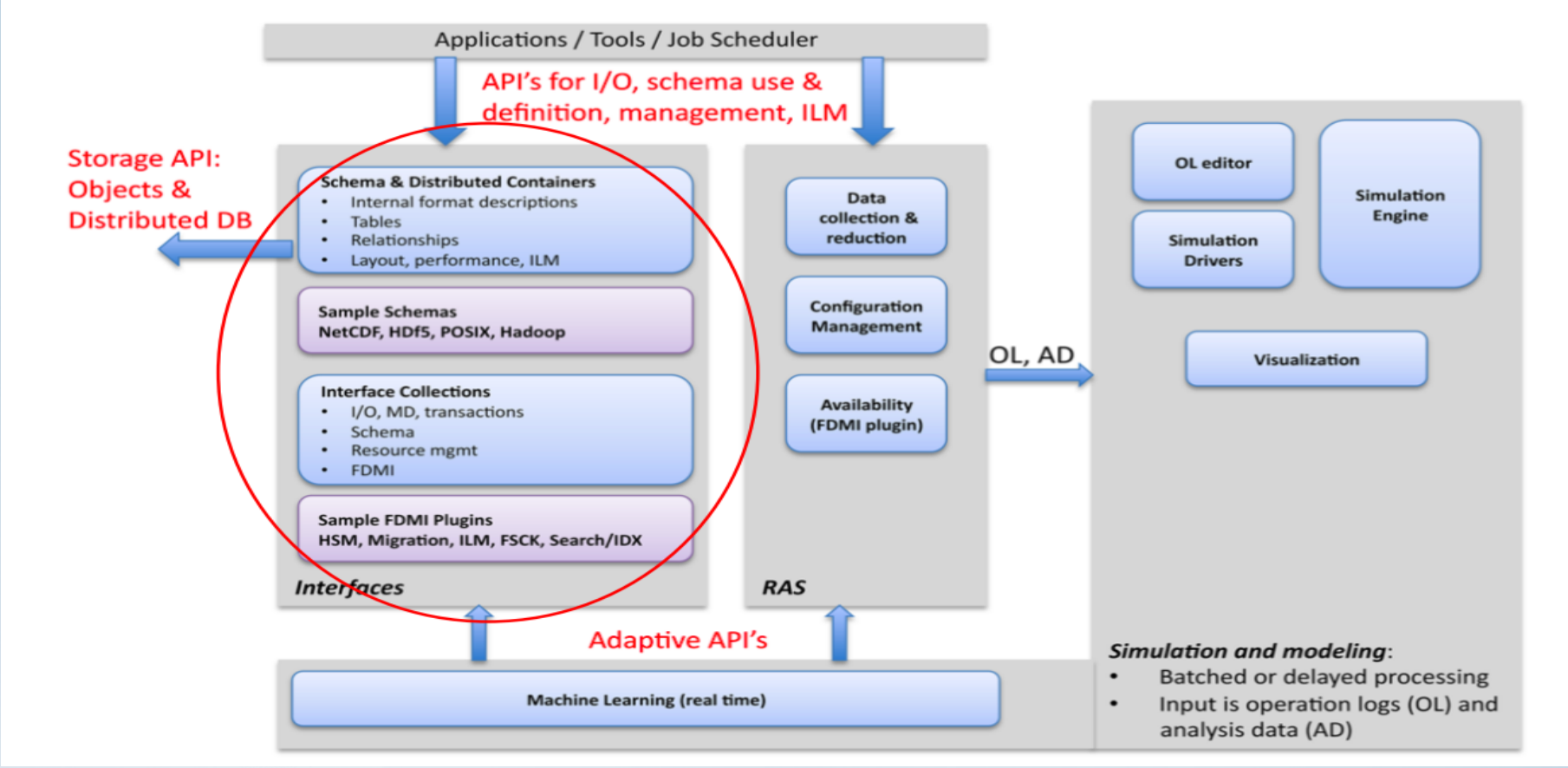
- Welcome and encouraged
- Especially for RED items

If this is a project you'd like to see funded, let your EOFS and OpenSFS representatives know!

E10 Update

xyratex

E10 Core Components



xyratex

E10 Core Components Update

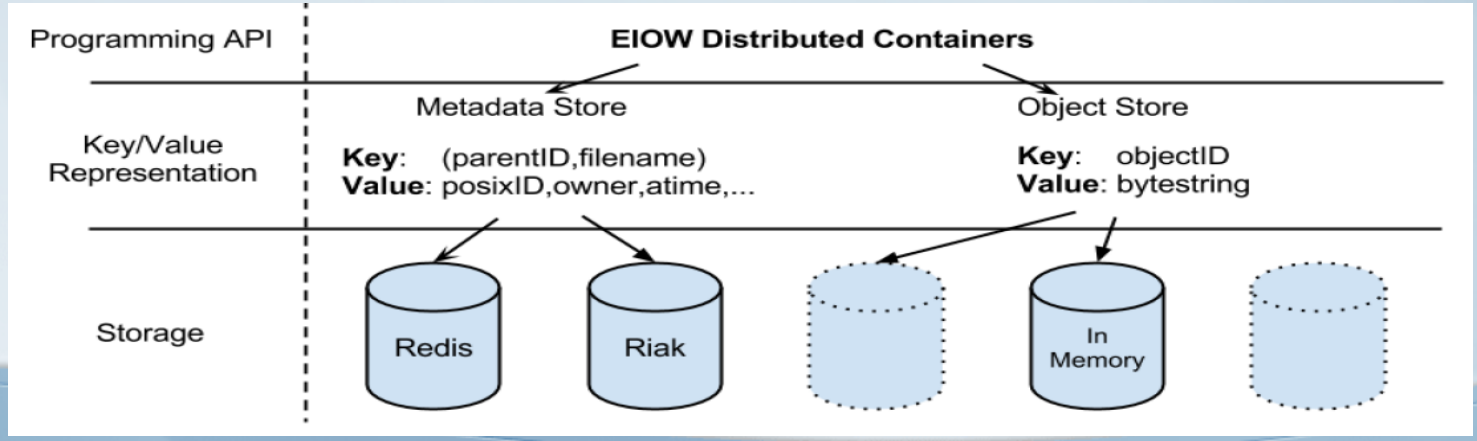
- **Implementation in Progress(prototyope builds)**

- "Schemas" and Distributed Containers
  - POSIX Schema being worked on
  - Use of Haskell Programming Language

- Storage Interface and Implementation
  - In-Memory for debugging/testing - for now

- Container Manager Implementation
  - Namespaces for distributed containers

Contact:
JonathanJouty/ParSci



xyratex

HA and API Updates

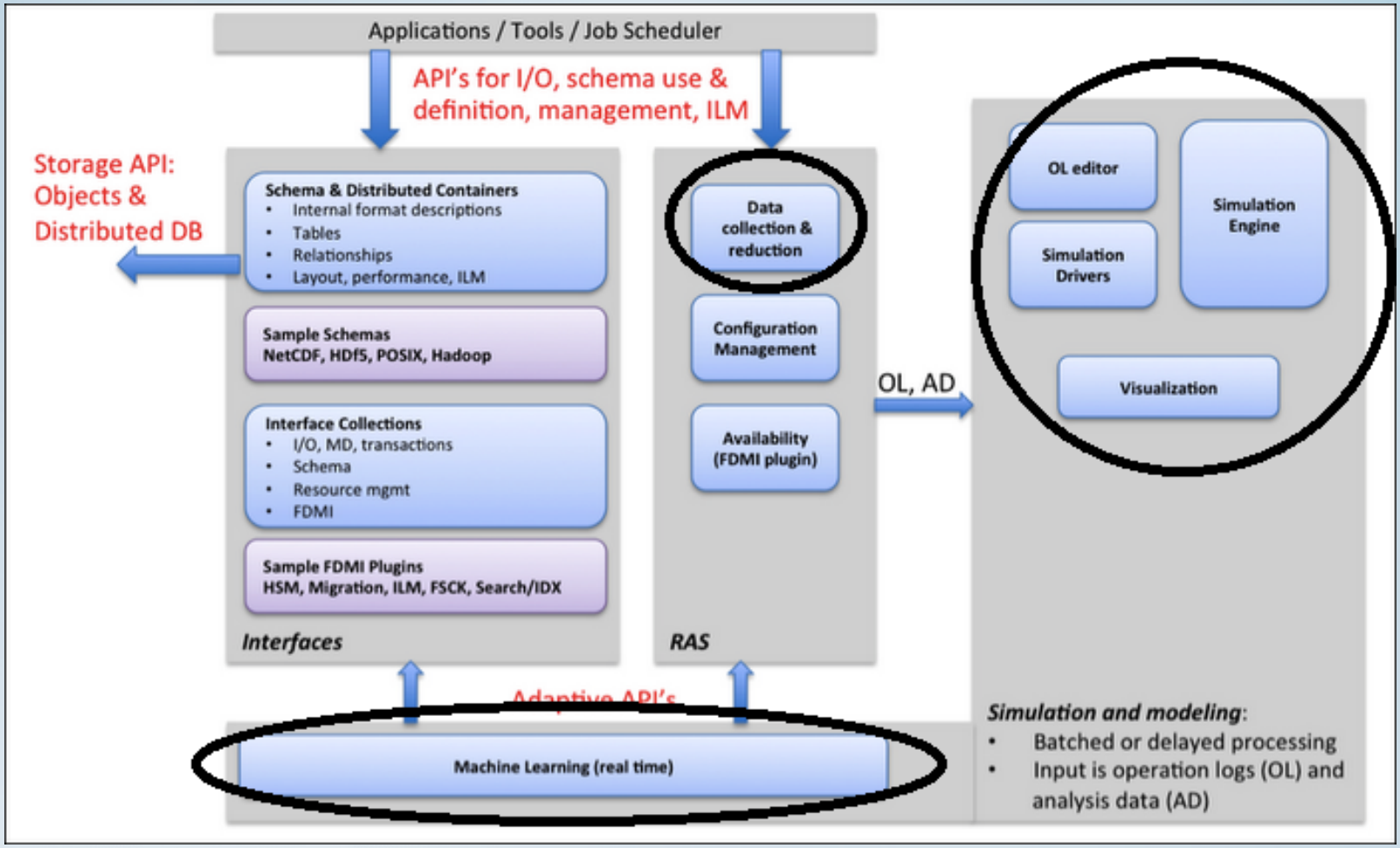
- **High Availability**
  - Highly available architecture for Infrastructure Nodes
    - Paxos type algorithms being worked out
      - Achieving consensus among nodes
      - Implementation in progress (Haskell)

Contact:
Matthew Boespflug/Parsci

- **File system interfaces for Exascale**
  - “Clovis” interface being discussed
    - Object store interface
      - Key-Value based Metadata
    - Concept of transactions
    - A “full” interface
      - Complete storage interface for E10
      - Concept of resource management and layouts

Contact:
Nikita Danilov/Xyratex

E10, Sim/Mod/RAS: Primary Components



E10 Simulation, Modelling and RAS: Primary goals

- **Observability: Always "On" Infrastructure Telemetry data at Exascale**
  - Needs specialised analytics solutions (100s of TBs/day of Telemetry data for reasonably large clusters)
  - Specialised Anomaly Detection and Root Cause Analysis methods
- **"What if" Predictive Capability at Exascale**
  - What if we provide a flash tier?
  - What if we introduce PCM(Phase Change Memory) in the mix?
  - What happens as we scale this architecture out?
- **Learning Engine -** learns from the above components and helps the infrastructure to adapt

E10 Sim/Mod/RAS: Data Collection, Reduction and Machine Learning

- **Current Activity for E10 through the SIOX project - Led by University of Hamburg**
  - Project involves data collection and Analysis of activity patterns and performance metrics targeted towards Exascale I/O
    - SIOX aims to provide fine grain system performance
    - SIOX aims to locate and diagnose problems
    - SIOX "learns" optimizations to be fed back into the Exascale I/O system
- **Current Status**
  - High Level architecture available for the different subcomponents of SIOX
  - Early prototypes under development utilizing basic SIOX libraries

Contact: Julian Kunkel at the University of Hamburg

E10 Sim/Mod/RAS: Simulation and Associated Components

- **Current Activity for E10 through the Exascale I/O simulation Framework - Led by Xyratex**
  - Project involves development of
    - Exascale I/O Simulation Engine
    - Operation Log Editors for editing Exascale I/O workloads
    - Simulation drivers that provide models of I/O hardware components at Exascale
- **Current Status**
  - High Level architecture available for the Exascale I/O Simulation Engine
    - Utilizes Queuing Frameworks

Contact: Sai Narasimhamurthy at Xyartex

Future Meetings

- Mid-October Asia events
- SC13 BOF

- Regular conference calls

Thank You

Meghan_McClelland@xyratex.com