Long Haul Hadoop

Steve Loughran



Recurrent theme

Issue	Title
HADOOP-3421	Requirements for a Resource Manager
HADOOP-4559	REST API for job/task statistics
MAPREDUCE-454	Service interface for Job submission
MAPREDUCE-445	Web service interface to the job tracker
HADOOP-5123	Ant tasks for job submission



In-datacentre: binary

- Hadoop RPC, AVRO, Apache Thrift...
- Space-efficient text and data
- Performance through efficient marshalling, writable re-use
- Brittle fails with odd messages
- Insecure caller is trusted!



Requirements

- Works through proxies and firewalls
- Robust against cluster upgrades
- Deploys tasks and sequences of tasks
- Monitor jobs, logs
- Pause, resume, kill jobs
- Only trusted users to submit, manage jobs





WS-*

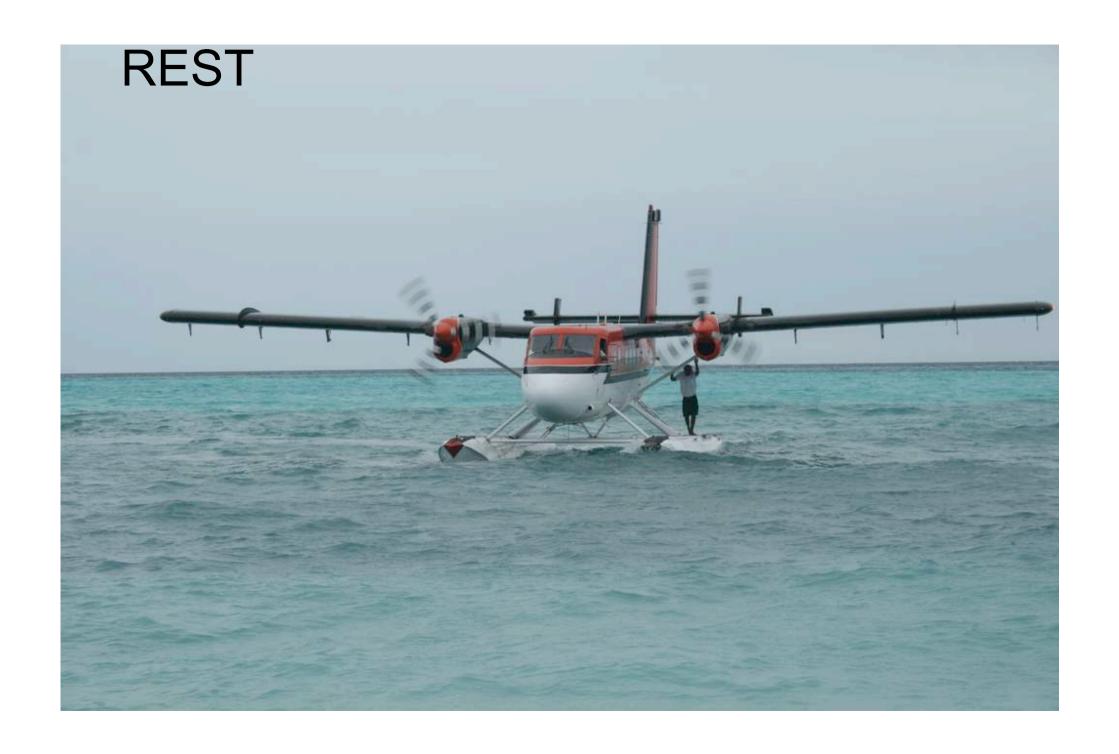
- ✓ Integrates with SOA story
- ✓ Integration with WS-* Management layer
- ✓ BPEL-ready
- Auto-generation of WSDL from server code
- Auto-generation of client code from WSDL
- ✓ Stable Apache stacks (Axis2, CXF)
- Security through HTTPS and WS-Security



WS-* Problems

- Generated WSDL is a maintenance mess
- Generated clients and server code brittle
- No stable WS-* Management layer (yet)
- WS-Security
- Little/no web browser integration
- Binary data tricky





Pure REST

- PUT and DELETE
- Cleaner, purer
- Restlet is best client (LGPL)
- Needs HTML5 browser for in-browser PUT

Model jobs as URLs you PUT; State through attributes you manipulate



HTTP POST

- Easy to use from web browser
- Include URLs in bugreps
- Security through HTTPS only
- HttpClient for client API
- Binary files through form/multipart

Have a job manager you POST to; URLs for each created job



JobTracker or Workflow?

- 1.HADOOP-5303: Oozie, Hadoop Workflow
- 2. Cascading
- 3.Pig
- 4.BPEL
- 5. Constraint-based languages

The REST model could be independent of the back-end



Status as of Aug 6 2009

- 1.Can deploy SmartFrog components and hence workflows or constrained models
- 2. Portlet engine runs in datacentre
- 3. No REST API
- 4. Need access to logs, rest of JobTracker
- 5. Issue: how best to configure the work?



