

Towards Persistent Scientific Data Archives

June 2006

Kevin Gamiel

Renaissance Computing Institute



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

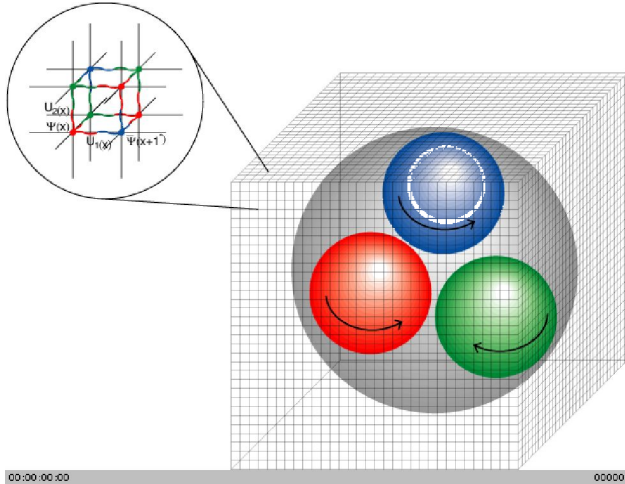


NC STATE UNIVERSITY

Background

- Renaissance Computing Institute (www.renci.org)
 - Collaborative venture of UNC, Duke, and NC State
 - Multidisciplinary, cyberinfrastructure
 - Oceanography, Meteorology, Biomedicine, HPC, Grid, ...
- Need for persistent access to arbitrary scientific data sets
- Need for discoverability within sets
- Technologies considered include THUMP
- Driving scenario: DOE SciDAC QCD data
 - Scientific Discovery through Advanced Computing
 - Quantum Chromodynamic quark model

Scientific Data Records



- **Software performance data records**
- **Instrumented QCD (MILC) code**
- **Collected across heterogeneous hardware platforms, tens to thousands of nodes**
- **Expressed in XML**

Source: Richard C. Brower & Robert Edwards



Commodity clusters
at FermiLab,
Jefferson Lab



IBM Bluegene/L
at Argonne
National Lab



Cray XT3 at
Oak Ridge
National Lab



SGI Altix at NCSA
and UNC-Chapel Hill

Anatomy of a Performance Record

Common Metadata

User info

Application Info

Platform Info

Environment Info

Profiling Data

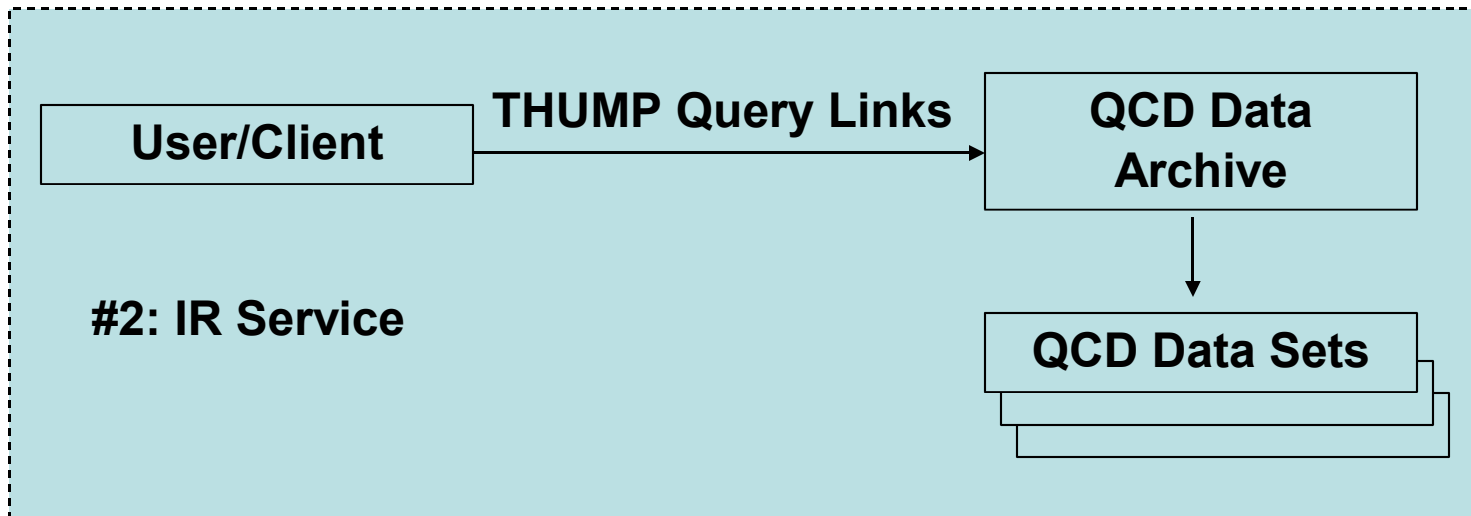
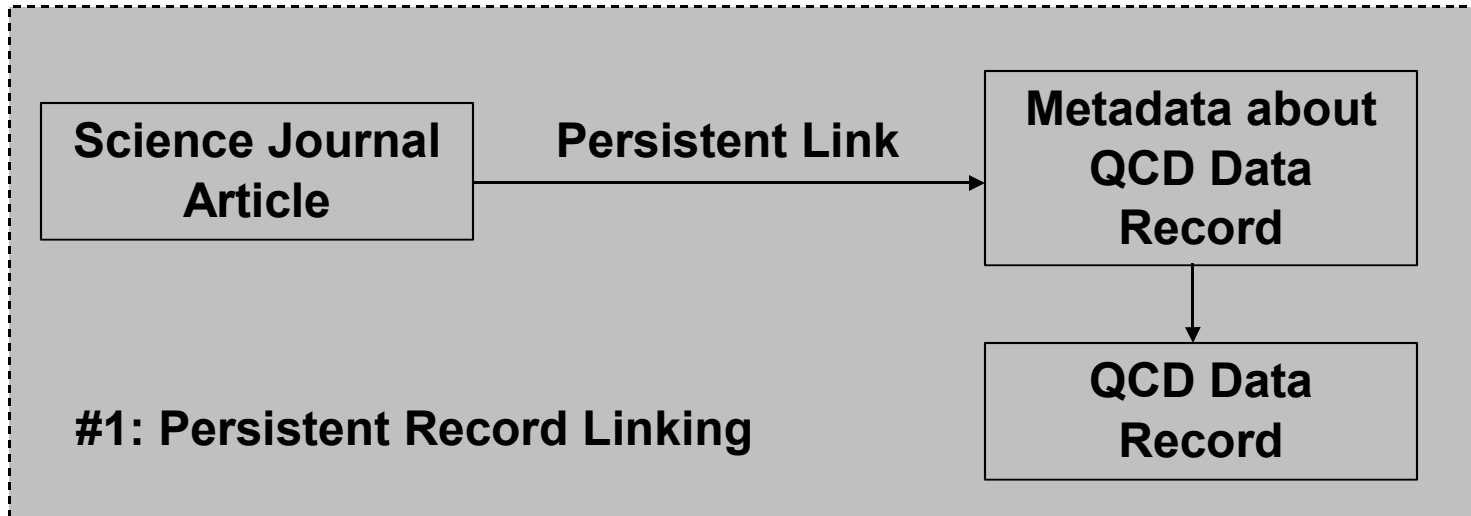
Instrumentation Point Events

Platform-Specific Data

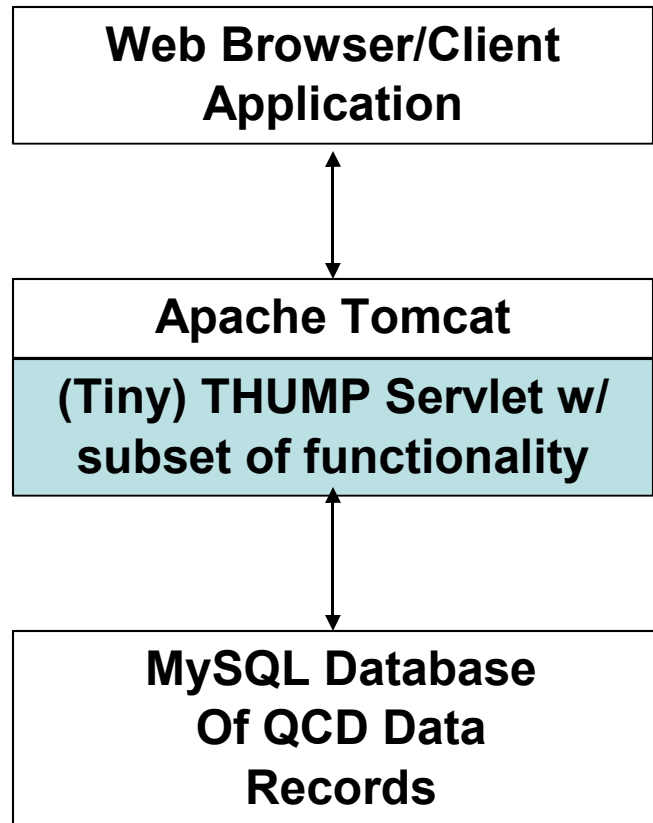
Hardware counters

Processor statistics

THUMP as a Solution – The Vision



Thump as a Solution - Implementation



Conclusions

- Classic simplicity is appreciated
 - RFC-in-a-day
- What worked well
 - Metadata retrieval service (esp. w/ ARKs)
 - ERC (electronic resource citation) simple, concise, human readable, easy to transcribe
 - Testing/Debugging via browser address field and/or telnet plus minimal keystrokes
- What could be improved
 - Consider standard URL encoding for value/attribute pairs
 - Some verb names not intuitive, e.g. “was”
- Future plans
 - RENCI will develop a complete THUMP implementation for continued evaluation