Streaming data for science and research

Paul Hubbard
Larry Miller
May 5th 2008
hubbard@sdsc.edu, ljmiller@sdsc.edu

Cyberinfrastructure Lab for Environmental Observing Systems
Science R&D
SDSC/UCSD
Talk outline

• Introduction and background
• Tech info
• Demo
• Code tutorial
• Where to get the code and docs
• Development roadmap
• Please ask questions at any time!
What is DataTurbine?

- DataTurbine is an open source, Java based network ring buffer for all sorts of data. You can use memory + disk for the ring and it runs on almost any JVM.
- Started life as a NASA telemetry project
- The basic division of work looks like this:
Why do you want a ring buffer?

- Problem: Stream data to multiple clients, with ability to rewind, recover from TCP errors and potentially handle different data rates.
- Multicast is nice but
  - You can’t assume its presence (often disabled at edge router by default)
  - Not easy to have different rates for different sinks
- One solution: Ring buffer, where span of ring is ‘large enough’ for each application.
DataTurbine tech details

- Ring buffers are per-source configurable with amounts of memory and disk. By using ratio of 10% or so, very large sets of data can be rapidly accessed. (Empirical result)
- Parent/child/*child routing topologies as well as simple mirrors
- Primary interface is Java API, but can also use
  - ActiveX on Win32
  - TCP/UDP proxy interface
  - WebDAV/HTTP (most operating systems) via Tomcat app
  - Java proxy with arbitrary interface/protocol
- The killer app is probably the ability to navigate data TiVo-style (scan through, replay fast/slow, play backwards, etc)
- Time synchronization - server/client, NTP necessary
More about DataTurbine

- Sources can have multiple channels with varied types - numeric (e.g. sensors), video, audio, text, binary blobs.
- We have a variety of sources and sinks: In-house, from the original author Creare and also community contributed.
- Can also use plugins for tightly-coupled computations such as image processing.
- Runs on J2ME, J2EE and 64-bit JVM as well. Extremely scalable.
A more complex example
Metadata in DataTurbine

- There are two kinds of metadata: static and time-varying
  - E.g. static: Per-channel units
  - Time-varying might be, for example, GPS position
- Static data is written with PutUserInfo and PutMime
  - cmap.PutUserInfo(idx, "units=G,scale=1,offset=0");
  - cmap.PutMime(idx, "application/octet-stream");
- Time-varying metadata is best put into its own channel
Design decisions and notes

- Timestamps are 64-bit, 32 each of time_t plus fractional second, stored UTC on the server by convention. RBNB doesn’t care.
- Timestamps can be user-generated (e.g. DAQ with accurate times) or API-calculated
- If computer clocks differ (e.g. server/client/sink) you’ll have no end of problems. You need NTP, especially on Windows which has SNTP by default.
  - [http://dataturbine.org/content/network-time-synchronization-and-rbnb-dataturbine](http://dataturbine.org/content/network-time-synchronization-and-rbnb-dataturbine)
- Every source defines its own cache & archive parameters
- Data is aggregated into ‘frames’ which are manually flushed
  - Allows tuning of packet size/update interval
More design notes

• Sources disappear at close unless explicitly detached
• We expect $RBNB_HOME to be defined
• Server exposes logfiles and internal metrics as normal DataTurbine channels
  • Hidden by default
  • You can use normal tools to monitor your servers
• New feature - set cache/archive by time
• Network + API = robust abstraction layer
Security and access control

- Server has IP-based access control now, very similar to hosts. {allow, deny}
- Server has currently-unused mechanisms for requiring username/passwords
- We’ve done some mailing-list surveys and this is currently a low priority.
Performance

- 65MB/sec on Solaris/sparc
- (T2000, 16GB, gig-e, JVM64)
- Student-authored test suite
- Also have MATLAB benchmarks
Macbook pro to T2000 over gig-ethernet - ~30MB/sec from MATLAB
...what about Antelope?

- Briefly, RBNB is neither as fast nor as complete as Antelope
  - Compiled vs Java
  - ORB protocol is probably more efficient on the wire
  - Custom database instead of binary chunks dumped to disk
  - Far fewer tools (admin, view, plot, extract, etc)
- We have beta code for orb-\(\rightarrow\)DataTurbine
- DataTurbine is free and open source (Apache 2.0 license)
- Choose your poison!
Who are we? NSF, SDCI and CLEOS

• In summer 2007, the CLEOS group at SDSC won a two-year NSF award under the SDCI (Software Development for Cyberinfrastructure Improvement) to work on DataTurbine
  • Move from closed-source to community-based open source (Apache 2.0 license)
  • Create and record metrics for performance, scalability
  • Port to 64-bit Java
  • Work with various communities to encourage use and dissemination
  • User workshop fall 2008
Viewing, browsing and analyzing data

• Getting data into DataTurbine is often the easy part. Once there, you need a good viewer that lets users interact with the data in ways that they find useful.

• There are many clients (sinks) as well as DataTurbine->SQL code, file writers, etc so you can use existing tools
  • Simple interfaces, import/export both lower the difficulty of using DataTurbine. We don’t want to be a one-stop-shop.
RDV is the Real-time Data Viewer, written by Jason Hanley at SUNY Buffalo for NEES. It’s plugin-based Java, handles time series, X vs Y, FFTs, audio, video, TiVO-style navigation, per-channel metadata, events and more.
More RDV
Deployments of DataTurbine

• I’ve included some screenshots of DataTurbine in use to give a flavor of current utilization.
CLEEOS/HPWREN deployment at Santa Margerita Ecological Reserve
Ecological reserve
NCHC (Taiwan)

- Kenting National Park and Yuan-Yang Lake, pictures from Fang-Pang Lin and Ebbe Strandell
Kenting (NCHC/Taiwan)
More Kenting
Insight Racing

- DARPA autonomous vehicle competition
- Insight is using DataTurbine for their vehicle video in their Lotus
- North Carolina State University, using multiple Axis 206 cameras, 30fps each
- http://www.insightracing.org/
NASA Dryden Flight Center

- Intelligent Network Data Server (INDS)
- Fusion of DataTurbine, Google Earth and live telemetry
- Instruments flown on ER-2 (U2) and DC-8
One more NASA
GWT-based web interface
Current device support

- **DAQ**
  - National Instruments (NI-DAQ, compact RIO) via Java proxy DaqToRbnb
  - Campbell Scientific: File-based, via LoggerNet, up to 1Hz tested. Download/parse/upload
  - Dataq Instruments (serial connect via C + DaqToRbnb)
  - Davis Instruments - planned, via Meteo
  - PUCK, Seabird

- **Video and still cameras**
  - Anything with motion JPEG via URL (AxisSource)
  - Still images via WebDAV

- **Accelerometers - ADXL202 and Apple laptop**
Live demo!

- Run DataTurbine
  - cd $RBNB_HOME/bin
  - java -jar rbnb.jar

- Now run source(s)
  - cd code/sms-rbnb
  - ant run

- Run RDV
  - cd code/RDV
  - ant run
  - /File/Connect/localhost:3333
Live demo: adding features

- **Add isight stills via WebDAV**
  - $RBNB_HOME/bin/Start_Webserver.sh
  - mount localhost:8080/RBNB
  - mkdir /Volumes/RBNB/iSight
  - cd /Volumes/RBNB/iSight
  - watch isightcapture -w 640 -h 480 -t jpg isight.jpg

- **Add PNGPlugin**
  - open http://localhost:8080/
  - Start PNG Plugin with defaults

- **View via web page**
Live demo: part three

• How about better video?
  • Kill isightcapture
  • Run ISightToRbnb
    – java org.nees.rbnb.ISightToRbnb -n ISSource

• What’s the server up to?
  • RDV/View/Show Hidden Channels
  • Plot metrics

• Event markers: timestamped metadata
• Channel from ISightToRbnb vanishes - discuss!
Live demo: part four

• Go to http://tinyurl.com/2mjog2
  • File/Connect {my IP}
• Parent/child routing
• Optional if interest and time
  • Admin interface
  • Plot.jar
  • Source.jar
• SMER data on niagara-stage.sdsc.edu
Future plans

• We have NSF (SDCI) funds to improve, extend and enhance DataTurbine over the next two years, and other funds to support a variety of deployments.

• Plans to
  • Add triggered video data (iQeye, Axis)
  • Web display
  • Collaborate as much as possible, with an eye towards building our community

• Major effort right now on better mirroring
Where to learn more

- Code, documentation, screenshots, developer mailing list, FAQ, Wiki and more are all available at http://dataturbine.org/
- We are very interested in developers, collaborators, users and in generally pushing the technology to new areas and capabilities.
- We have dev/stage/prod turbines for to try and use on niagara(-dev, -stage).sdsc.edu
- Thank you!