WP8: FlexBuff

Ari Mujunen, Tomi Salminen
Aalto University Metsähovi Radio Observatory

EVN TOG Meeting – 2012 June 27 – Onsala Space Observatory, Sweden

Research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement n° RI-261525. This presentation reflects only the author's views.
The European Union is not liable for any use that may be made of the information contained therein.

NEXPReS

WP8: FlexBuff

- The full title in the “Description of Work” is really, really long:
  - “WP8: Provisioning High-Bandwidth, High-Capacity Networked Storage on Demand”
- Thus introducing a new acronym for these “flexible buffers”
  - Before the unofficial one “AriBox” gets adopted irrevocably… :-)

Primary Objective

- Be able to reliably record a local high-speed UDP packet stream onto local buffer disks
  - dBBC/FiLa10G, Roach1/2 variants, iBOB...
- Allow simultaneous long-distance remote read access, for correlation processing
  - Long-distance & high-speed implies UDP

Key Implementation Details

- Don't waste CPU by making multiple memcpy() copies of 10Gbps rate data
  - Implies O_DIRECT or libaio asynch disk I/O
  - UDP packet processing just once
- Ensure disks get enough work per every r/w call (& accompanying seek)
  - About disk hw cacheful / one trackful, tens of megabytes
- Keep all disks equally spinning
  - Easiest with every disk having its own regular filesystem
    - => Unlike raid0, loss of one disks only loses part of data
  - Use (most of the) main memory to allow each disk write its large sequential chunk of continuous data

vlbi-streamer

- Available at:
  - http://code.google.com/p/vlbi-streamer/
- Recording local UDP streams
  - Concurrent interleaved reading of the same disks
- Shows that initial architectural decisions were correct by its performance
  - Large enough direct/asynch I/O blocks per each disk

Recent vlbi-streamer Test Results

- Local UDP streaming performance tests
  - Wirespeed 10GE
  - Long (30min) tests show writing at max wire speed and 0 ploss
  - Writing to 34 disks /wo net
  - Architecture can handle 40Gbps; always >30Gbps. Close to HW-limit.
  - Using INAF's hw RAID system (/wo net)
  - Works reasonably well even there, 10-14Gbps. Lower performance is due to HD access granularity (~32MB writes). Speed increased by 4Gbps with larger writes (256MB-512MB).
- Long-distance UDP streaming tests
  - NORDUnet does not have a lightpath up anymore, so Jb <-> Mh tests on hold

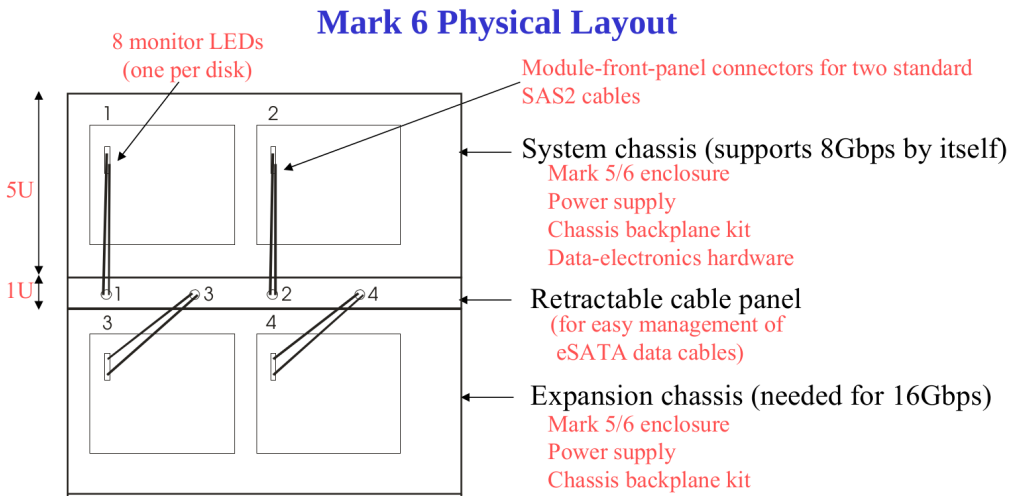
Recent Haystack Developments: Mark6

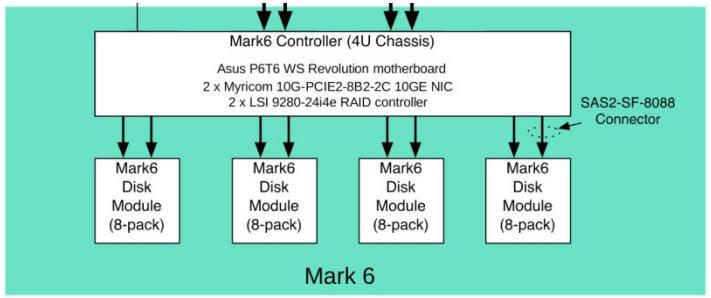
- Essentially the same hardware & Linux except:
  - Except hard disks are in external enclosures
  - Physically nearly identical to old Mark5 8-disk packs
  - Two external MiniSAS connectors added to the front panel
- Same software can run on WP8 & Mark6 hw
  - Current Mark6 software is Mark5C-style raw Ethernet packet capture onto raid0 volumes
  - http://vdas.org/
  - And WP8 hw can have controllers with ext MiniSAS

FlexBuff / Mark6

- Just different “packaging” of disks
  - Fixed vs. shippable

5U, 36 disks





Mark6 figures: “Mark6 data system DiFX presentation”, A.Whitney, D.Lapsley, 2011.12.05, http://www.haystack.edu/tech/vlbi/mark6/mark6_memos/04-2011.12.05_Mark6_data_system-DiFX_mtg-Haystack.pdf