THE UBC vn BEOWULF CLUSTER

- CFI Proposal & Funding

- Construction

- Applications

- Future Plans

Matthew W. Choptuik, UBC & CIAR
TRIUMF Farming Mini-Workshop, Vancouver BC, February 28, 2000

Supported by NSERC, CIAR, CFI and NSF PHY9722068

Standard prefix: laplace.physics.ubc.ca:/People/matt/

The Canadian Foundation For Innovation

www.innovation.ca

The CFI was established by the federal government with an up-front investment of $800 million. This principal amount and accrued interest will enable the Foundation to contribute, on average, about $180 million annually over five years to research infrastructure projects. The CFI targets its investment at key needs in the areas of health, environment, science, and engineering. The Foundation operates on the principle that its investments are made in partnership with the private and voluntary sectors as well as with provincial governments. The Foundation contributes 40% of total eligible project costs. On this basis, funding for the total investment by the Foundation and its partners should exceed $2 billion.

- Several programs, including *On-going New Opportunities*
  - Eligibility: First tenure track position in Canada.
  - HPC Potential at UBC from CFI: $\sim$ 750–1000 K / yr

The UBC vn PIII/Linux Cluster

Doc/VN/index.html

- 280K CFI On-going New Opps. App., 4/29/99 (UBC)
  Doc/CFI.april99/index.html
  - Affleck (Phys. & Astro.)
  - Ascher (Comp. Sc.)
  - Choptuik* (Phys. & Astro.)
  - Patey* (Chem.)
  - Salcudean* (Mech. Eng.)
  - Thachuk* (Chem.)
  - Unruh (Phys. & Astro.)

- Patterned after Patey/Thachuk's machine (currently 23 compute nodes and one front-end, roughly half done), asks for
  - 64 $\times$ Dual 450 Mhz PIII/512 Mb/10 Gb (no CD ROM, keyboard, mouse, monitor) "compute nodes" 220K
  - 2 $\times$ Dual 450 Mhz PIII/512 Mb with additional peripherals "front-end nodes" 10K
  - 1 $\times$ HP-4000M Switch with 4 expansion modules $\rightarrow$ 72 (!) 100FDX ports (3.6 Gb/s back-plane) 7K
  - 13 (!) $\times$ APC Smart-UPS 1400 14K

The UBC vn PIII/Linux Cluster

- 650K CFI On-going New Opps. App., 9/15/99 (CFI)
  `Doc/CFI/index.html`
  - Affleck (Phys. & Astro.)
  - Ascher (Comp. Sc.)
  - Bushe* (Mech. Eng.)
  - Choptuik* (Phys. & Astro.)
  - Patey* (Chem.)
  - Salcudean* (Mech. Eng.)
  - Thachuk* (Chem.)
  - Unruh (Phys. & Astro.)

- ASKS FOR "Cluster 1" *AND*

- "Cluster 2" (focus on coarse-grained parallelism)
  - 48 $\times$ Single 600Mhz Alpha/2 Mb/256 Mb/10 Gb 230K
  - Myrinet (1000 Mb) Switch solution 32K
  - 8 $\times$ APC Smart-UPS 1400 9K

- CFI funding for full amount (40% of 650K) awarded in December 1999, matching BCKDF funding (40%) expected (but still not secured), UBC will fund remaining 20%

The UBC vn PIII/Linux Cluster

- 280K for vn advanced against future CFI funding 8/27/99
- 9/99–10/99 spent evaluating machines, finding good location, setting up bid details with Purchasing
- Request for bid sent out 10/7 with closing date 11/2, equipment to be delivered 16 nodes per week
- Vendors: *Varsity*, *UBC Bookstore, AE*
- WHAT WE PURCHASED
  - 64 compute: 2 x 450Mhz PIII/512 Mb/10 Gb IDE 180k (Memory: 44% of total compute node cost)
  - 3 front-ends: 2 x 450Mhz PIII/512 Mb/34 Gb SCSI 20K
  - 1 $\times$ HP-4000M Switch: 7K
  - 4 $\times$ APC Matrix 3000M with 8 PDUs: 19K
- Total hardware expenditures: $<$ 250K
- Computer room (Klinck/Old CS) annual "rent": 7K
  Currently split three ways: Klawe, Bushe, Choptuik

The UBC `vn` PIII/Linux Cluster

- Assembly & Software Installation Team
  - Jason Ventrella
  - Inaki Olabarrieta
  - Choptuik
  - Unruh
- At vendor (near UBC)
  - BIOS settings
  - “Everything” (!) install of Mandrake 6.1
  - Network configuration including IP address assignment
- At our site (Main Machine Room, Old CS Bldg)
  - Plug node in, attach to network, power up
  - Secondary software installation (remote)

The UBC vn PIII/Linux Cluster

- Hardware shake-down, $\sim$ 6 problem nodes identified in first few weeks of operation: power supplies, bad memory, second processor not recognized
  - Power supplies
  - Bad memory modules
  - Second processor not recognized
- No hardware problems since then.
- Software problems
  - Driver for Intel network card required updating
    Rebuild of kernel required
  - Intermittent problems with MPI (mpich)
    Unresolved, currently re-boot offending nodes
- About 50 "incidents" / "crashes" so far, but impact on users is minimal—typically bad node(s(s)) reported, user moves onto other node(s).
- Entire cluster has never been taken down.

vn.physics.ubc.ca: First 16 compute nodes & 3 front-ends

vn.physics.ubc.ca: Back-end View

The UBC vn PIII/Linux Cluster

Sample Applications

- "shell-level" parallelism
  - Ethan Honda (UT Austin grad stud): detailed parameter space survey of "oscillons" (typically 40 + processes)
  - Roman Petryk (UBC grad stud): quantum gravity inspired calculations (typically 40 + processes)

- MPI-based parallelism
  - Luis Lehner (UT Austin postdoc), Mijan Huq (Penn State postdoc): 3D black hole calculations
    (81 x 81 x 81 spends 11
  - Roman Baranowski, UBC Chemistry postdoc (MD Simulations)
  - Lothar Buchmann, TRIUMF Research Scientist (Nuclear Physics)

The UBC vn PIII/Linux Cluster

The anarchy queueing system

```
vnfe1 % ruptime | grep -v down | grep -v vnfe | sort -n +6

vn10 up 9+11:34, 0 users, load 0.00, 0.00, 0.00
vn11 up 9+11:34, 0 users, load 0.00, 0.00, 0.00
vn13 up 9+11:34, 0 users, load 0.00, 0.00, 0.00
vn15 up 9+11:31, 0 users, load 0.00, 0.00, 0.00
vn20 up 9+11:32, 0 users, load 0.00, 0.00, 0.00
vn21 up 9+11:32, 0 users, load 0.00, 0.00, 0.00
vn22 up 9+11:32, 0 users, load 0.00, 0.00, 0.00
vn23 up 9+11:28, 0 users, load 0.00, 0.00, 0.00
vn24 up 9+11:28, 0 users, load 0.00, 0.00, 0.00
vn26 up 9+11:29, 0 users, load 0.00, 0.00, 0.00
vn35 up 9+11:27, 0 users, load 0.00, 0.00, 0.00
vn39 up 9+11:27, 0 users, load 0.00, 0.00, 0.00
vn40 up 9+11:28, 0 users, load 0.00, 0.00, 0.00
vn41 up 9+11:28, 0 users, load 0.00, 0.00, 0.00
vn42 up 9+11:28, 0 users, load 0.00, 0.00, 0.00
vn43 up 4+01:51, 0 users, load 0.00, 0.00, 0.00
vn44 up 4+22:16, 0 users, load 0.00, 0.00, 0.00
vn8 up 9+11:34, 0 users, load 0.00, 0.00, 0.00
vn9 up 9+11:34, 0 users, load 0.00, 0.00, 0.00
vn33 up 9+11:26, 0 users, load 0.97, 0.91, 0.82
vn38 up 8+17:48, 0 users, load 1.82, 1.91, 1.89
                    .
                    .
                    .
vn53 up 4+21:31, 0 users, load 2.27, 2.20, 2.08
```

FUTURE PLANS

- Second cluster from remaining CFI/BCKDF/UBC funding ($\sim$ 300K)
  - Projected installation: 11/00
  - 48 $\times$ Single 600Mhz Alpha/2 Mb/256 Mb/10 Gb
  - 1000 Mb interconnect

- Plan to submit UBC-wide (BC wide??) proposal for last CFI "Major Facility" competition.
  - Will probably ask for 2000K+

- $<$ 1000K / yr CONTINUOUS for hardware gets UBC world-competitive in SOME (MANY?) simulation-based areas

  500K / yr keeps us in the game

- If you build it, they (talented, motivated grad students, post-docs, young profs, $\cdots$) will come and run the machines (?!)

- Possible participation/partnering with one of NSF supercomputing centres?