Experiences Tuning a J90 for Engineering Applications

Sharan Kalwani

shan@cray.com

CUG 1997, San Jose

Motivation

- Why do sites choose a J90?
  - Cost is usually the primary consideration
  - Computational needs satisfied by J90
  - Starting point for future growth
  - Resources are also a major factor

Motivation (contd...)

- Field experience revealed:
  - Several sites out there (250+)
  - Site profiles were all different , eg:
    - Human resources often limited
    - Expertise scarce
    - Learning curve was steep
    - Balancing act of system versus “real” work

Benefits

- Share useful tips and techniques
- Gain insight into other possible scenarios
- Learn what works
- Learn what may NOT work

Caveat Emptor

- Target Audience
  - Engineering Applications
    - Computational Fluid Dynamics (CFD)
      - STAR-CD, KIVA-2, CHAD
    - Metal Deformation Codes
      - (LS-DYNA)
    - Finite Element Analysis (FEA)
      - NASTRAN, etc...

Typical J90 Environment

- Usually small systems:
  - 4 to 8 CPUs
  - 64 Mwords
  - Single IOS
  - SCSI disks (sometimes also IPI disks)
  - small disk farm
  - no large online tapes (4 or 8mm only)

J90 Environment (contd)

- Staff at sites:
  - single, but multi-tasking person :-)
  - has a real job in addition to Cray work
  - usually sophisticated end user, but
  - not a “pure” systems type

Tuning Regions

- UNICOS Kernel
- Memory and
- Process Scheduling
- Disk Drive Strategy

Tuning: J90 Kernel Arena

- UNICOS Kernel
  - Impacts the overall running of the system
  - Even without source, some tuning possible
  - Requires however a few iterations to get things right
  - Often affected by the applications mix

J90 Kernel Arena

- UNICOS Kernel:
  - – mcache (or NBUFs) often an easy target
  - – default was 5% of available memory:
    - typically 6000 buffers for a 64 MW system
    - usually configured as 4096 on stock systems
  - – Can be set to 2048 for known mix

J90 Kernel

- UNICOS Kernel
  - mcache or NBUF can be done at UNICOS run time via param file changes
  - dramatic effect on job mix run
  - careful when tweaking buffers, too low a value chokes many a system.

J90 Kernel

- UNICOS Kernel:
  - – Network Buffers or TCP/IP space
  - – MBUFS typically set at 1800
  - – can be varied depending upon
    - network media type (eg Ethernet, FDDI...)
    - application requirements
    - Usual range 1200 to over 18000!
  - – change “TCP_NMBSPACE” in param file

J90 Kernel

- UNICOS Kernel:
  - Network send and receive space:
  - Use netvar
  - Depends a greta deal on netw patterns
- NFS:
  - Use large “rsize” and “wsize” on mounts
  - 4096 default, can be upto 32768.

J90 Kernel

- UNICOS Kernel:
  - To improve certain types of I/O
    - look at sar reports, eg for listio (sar -z)
    - bumping of physical I/O buffers beneficial
    - not easy to do, requires kernel rebuild
    - typically set at 200, can be increased to 400
  - I/O tables eg LDDMAX, PDDMAX,etc...

J90 Kernel

- Net Gain:
  - Reduced kernel size,
  - Reduced table size,
  - Larger Available Memory for jobs

J90 Scheduling

- Memory and Process Scheduling:
  - Major difference experienced via
    - nschedv command
  - Default settings via
    - /etc/nschedv.day and schedv.nite scripts
    - usually conservative

J90 Scheduling

- On a 64 Mw system, nschedv sets the
  - hog pool to 100000 clicks (or 48.8 Mw)
- Due to kernel “shaving”, this pool effectively rose to
  - 122700 clicks (or 59.9 Mw!)
- Use extra memory for
  - jobs
  - ldcache, etc...

J90 Scheduling

- Since memory is the prime resource:
  - memory based NQS setup is important,
  - stock scripts need always be modified
  - NQS global memory value must be chosen carefully
  - usually 1.7 to 2x of available memory

Disk Strategy

- Memory and Disk Strategy:
  - NQS memory oversubscription limited to a small 1.x factor
  - stock swap is set to 1 or 2 disks
- Swap speeds limited to SCSI disks
- Answer:
  - Stripe disks!!

Disk Strategy

- Striping Swap area:
  - helps overcome slow SCSI speeds
  - 3 to 4.5 Mbytes can be striped 2x or 3x
  - to effectively 10+ Mbytes/sec
  - useful even for IPI based disks
  - careful to separate swap disks from busy disks

Disk Strategy

- Striping Swap (contd):
  - I/O channel bandwidth is 50 Mbytes/sec
  - Effective to add more IOS channels
  - Initial swap setup costly, but drops thereafter
  - Also useful for checkpoint images via NQS

Overview

- The big picture:
  - J90 platforms can be tuned
  - Needs some effort to profile and effect changes
- Lessons from “bigger iron” can be also applied

Future

- Add lessons learned from other J90 configurations
- J90se and Scalable I/O systems

Where to get more information

- Training sessions:
  - UNICOS Performance & Tuning
- Books, articles, electronic sources:
  - UNICOS Tuning Guide, SR-2099
  - UNICOS System Admin Guide, SG2311
- Other sources:
  - FAQ in the works