z/TPF V1.1

TPF Users Group – Fall 2012

Title:
Performance and Reliability Topics

Robert Blackburn Ph.D.

AIM Enterprise Platform Software
IBM z/Transaction Processing Facility Enterprise Edition 1.1.0

Contents

- **Channel Redrive**
- **Local IPTE**
- **Memory Effects**
- **Dump Buffer**
- **CPU Measurement Facility**

Channel Redrive(CR) evolving value

- **Existed to preserve SAP(IOP) capacity**
  - All redrives done in channel engine
    - Total channel capacity very large compared to relatively scarce SAP power
  - Have measured up to 5 redrives per SSCH in customer workloads
  - Could have DEV,CU or switch busy
- **HW evolution has significantly lowered the value of CR**
  - CU queueing eliminates CU and DEV busy
  - FICON packet design means no switch port busy
  - In pure FICON environment there are 0 redrives

Channel Redrive limitations

- **With CR the SAP places the IO on a Round Robin(RR) selected channel**
  - We needed a minimal cost selection algorithm to maximize SAP throughput
- **RR works well with roughly equal server capacity**
  - If one path is weaker due to poor configuration or HW failure then RR will significantly overload that path
- **Standard IOP path selection will do a better job in optimizing unbalanced IO configurations**
  - They work with some function of response time

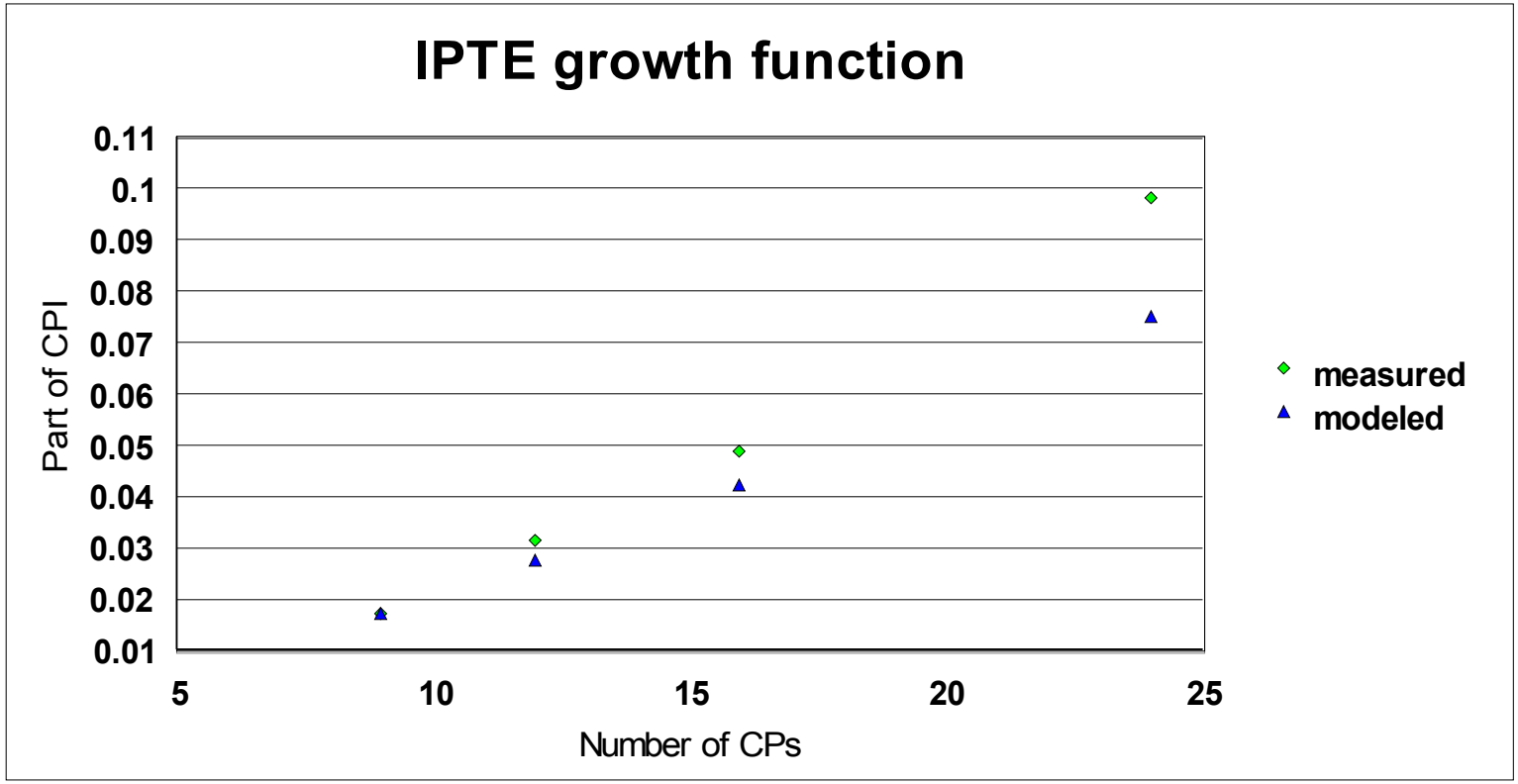
Customer should investigate running with no CR

- **With mixture of FICON/ESCON switch busies will still exist**
  - Examine SAP utilization for sufficient capacity
- **With mixture of CR and no CR CECs the CR CECs may see a slight reduction in IO service time compared to non CR CECs**
  - Have seen 7% delta in production
    - no CEC lockouts expected
    - essentially a sort of priority queue
  - Any sort of imbalance would make this delta vanish
  - All CECs no CR == all CECs CR
    - DASD response time

Invalidate Page Table Entry (IPTE)

- **TPF drives a fairly high rate of IPTEs mainly through Copy on Write**
- **IPTE has each CP update its TLB**
- **Key point------- CP issuing IPTE waits for all other N-1 CPs to update their TLBs**
  - Assume time to update TLB is random variable with exponential distribution
  - Maximum of N exponentials grows like (1+1/2+1/3 …)-> ln(N)
  - Total wait per CP grows roughly as N ln(N)

EC12 has 101 CPs for customers(120 total- 19 for SAP,Spare,Reserved)



Local IPTE(HW Change Requested by TPF Lab)

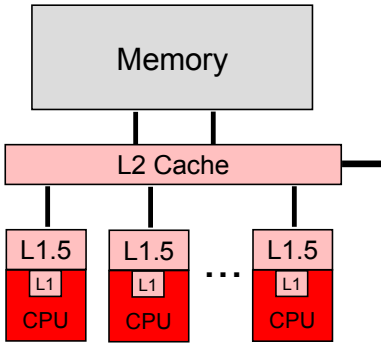
- **First available on the EC12 machine**
- **Local IPTE only updates TLB of issuing CP**
  - Other CPs unaffected and keep running productively
- **TPF now keeps history of CPs an ECB has executed on**
- **TPF makes certain the ECB has the latest copy of static data with every CP it runs on**
- **Can hold IPTE performance cost to a constant small level**
  - Greatly increases MP efficiency

Memory effects

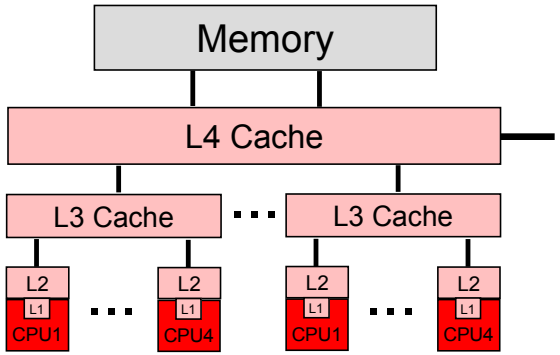
- **EC12 has**
  - 4 levels of cache
  - 2 levels of TLB
- **Increased cache structure needed for continued growth in total CEC MIPS**
  - Memory has been getting faster
  - CPU growing much faster

z196 versus z10 hardware comparison

- z10 EC
  - CPU
    - 4.4 GHz
  - Caches
    - L1 private 64k i, 128k d
    - L1.5 private 3 MB
    - L2 shared 48 MB / book
    - book interconnect: star



- z196
  - CPU
    - 5.2 GHz
    - Out-Of-Order execution
  - Caches
    - L1 private 64k i, 128k d
    - L2 private 1.5 MB
    - L3 shared 24 MB / chip
    - L4 shared 192 MB / book
    - book interconnect: star

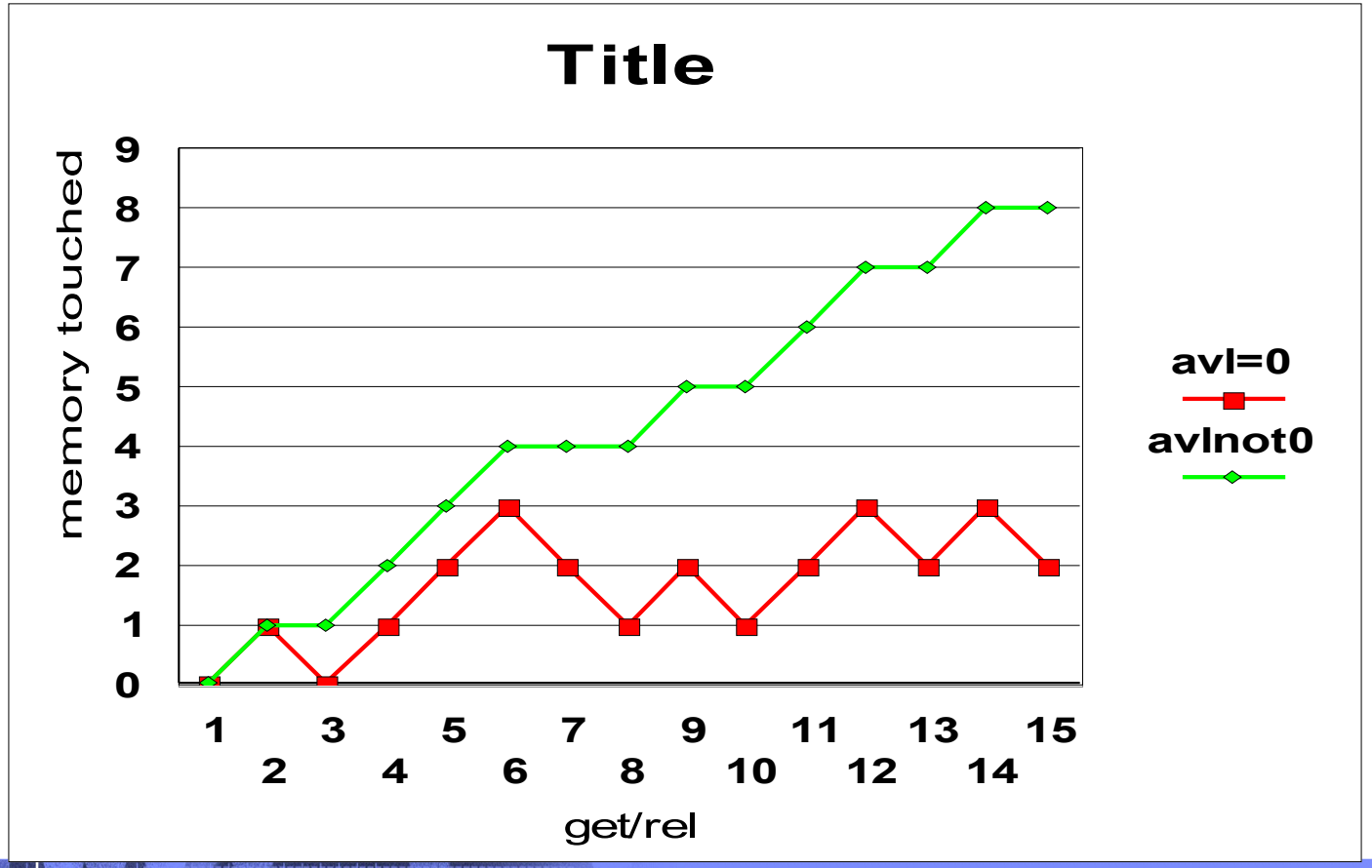


Memory Footprint Effects

- **When we think of performance often we think of instruction pathlength**
- **55 vs 300 trace elements in Lab experiments**
  - Suddenly mills/msg varied by 5-7%
  - Memory footprint per ECB increased
  - The instructions executed were identical

Avl=0 Memory Effect

- **Various cust---3 to 5% gain**
- **Debugging loss—not sig**



Dump Buffer can significantly improve TPF system availability

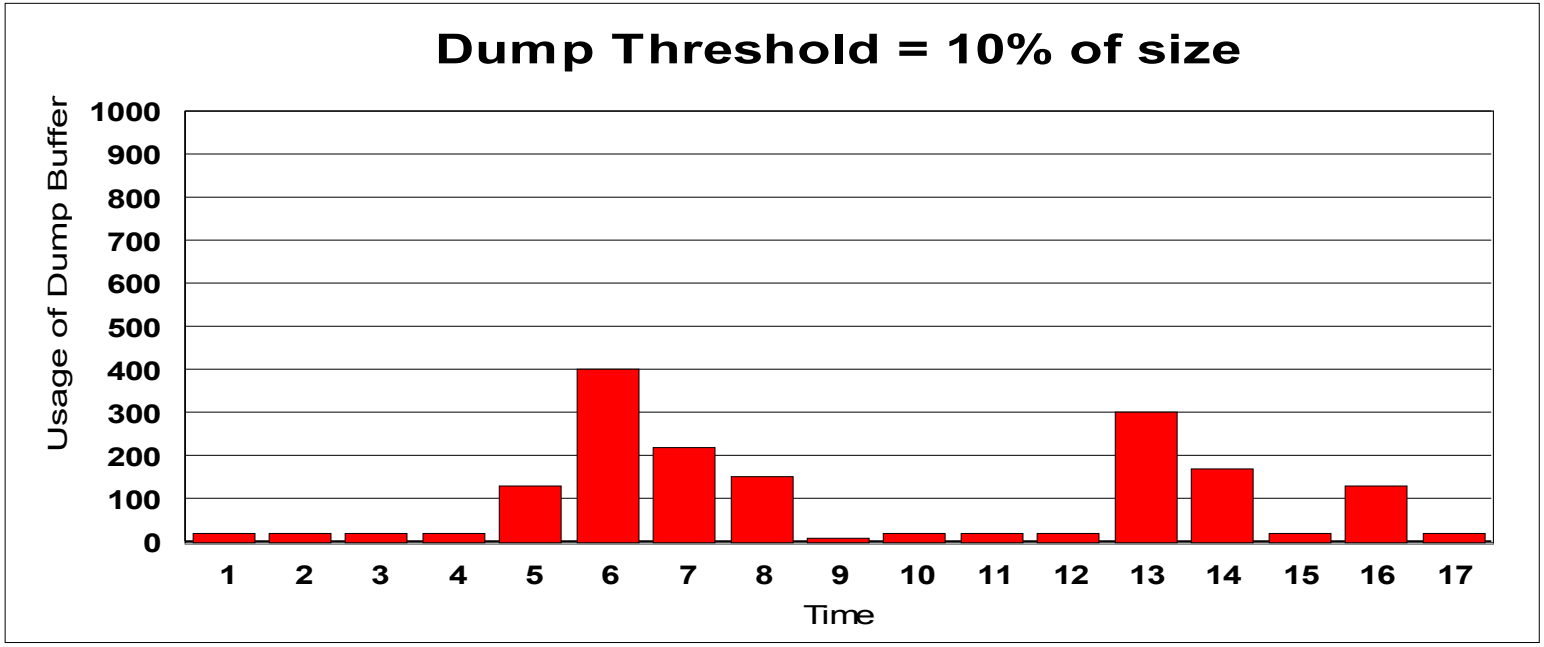
- **When dump to tape the size had to be carefully controlled**
  - 300M at 10M/sec => 30 seconds of CEC dead time
- **Rough time analysis for memory dump**
  - 256 line at 1000 cycles/line
  - 300M => ~1.2E9 cycles
  - Cycle time of .2ns => 1/4 second wait

How much dump space to define

- **Single LPAR in CEC then probably extra memory**
  - Dump buffer doesn't affect performance
  - VFA ~ 1G is a good rule of thumb
- **Too much space – extra memory cost**
- **Too little space – large probability that some dumps will go to tape**
  - View this as an outage

Dump Buffer data collection

- **Peaks over threshold**
- **Recorded values are 400M,300M and 130M**



Dump Buffer Sizing

- **Detailed customer dump pattern knowledge**
  - OPR and CTL dumps are ones of interest
  - Basically frequency x size with time correlation
    - For example
      - Mean rate of 30 dumps per hour
      - maximum of 3 CTLs in a 5 second period
      - max size < 300M
- **Stochastic approach**
  - Above had many not so easily checked assumptions
  - Record 2 to 3 months of dump buffer usage data
  - Feed that into and use Extreme Value Theory
    - 40% of Netherlands below sea level
    - 111 years of storm data
    - Government demand balancing cost and safety
      - Dikes built so Probability (overflow dike in a year) < 1/10000
    - Only 100+ years of data but yet can estimate an extreme quantile
  - Lab is willing to assist in this analysis

CPU Measurement Facility

- **First available on z10**
- **Absolutely critical for complete performance analysis**
- **Has been run at 2 customer sites – no problems**
  - From impact/risk view
    - think of it like running data collection
- **Lab is building HW usage profiles for**
  - RES,GDS,Finance,Rail …
  - Feed into future
    - processor development
    - TPF development

Trademarks

- IBM is a trademark of International Business Machines Corporation in the United States, other countries, or both.
- Apache is a trademark of The Apache Software Foundation.
- Other company, product, or service names may be trademarks or service marks of others.

- Notes
- Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.
- All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
- This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

- All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

- Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

- Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

- This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.