CREDIT SUISSE

Linux Kernel Summit 2007 September 5-6, 2007

Enterprise Panel

Date: September 6th, 2007

Produced by: “Head” Bubba (with contributions from others in various sectors)
IT Research & Development
contact: Head.Bubba@credit-suisse.com
Head.Bubba@ieee.org

Linux Kernel Summit
Credit Suisse

- Leading Bank
  - Approx: 45,000 employees, 50 countries
  - HQ in Zurich
  - Linux use: (example)
    - Trading Floor applications (Investment Banking/Fixed Income)
    - Database
    - Etc…
  - www.credit-suisse.com
- IT Research and Development
  - We look/develop/partner on the creation of new technology to solve our future business needs
  - We have a lab in which we test new innovative technology with real applications
- Want to interact with the kernel community directly
  - Desire to participate in kernel development
  - Willing to provide test applications that mimic application performance
  - Willing to test real applications on the latest kernel
  - Want to work out details for develop/enhance/contributing code
  - BUT – CS developers feel kernel developers disconnected from Customer Needs!

Linux Kernel Summit
Key Issues Today

- (1) Kernel Scheduler
  - Changes can have dramatic & detrimental impacts on applications
  - Applications are highly distributed; small changes to scheduler can impact latency and performance throughout a distributed system, including network and storage impacts
  - User level control of process priority and scheduling is becoming critical

- (2) Real Time Linux
  - Systems need low/predictable latencies
  - Performance benefits of not being interrupted
    - One example: 40% performance improvement was seen using a real time kernel (by isolating key threads & preventing interrupts to them); more work is being done
    - Existing isolcpu interface does not seem to be enough
      - Requires boot time isolation (not dynamic), and does not provide capability to schedule more than one thread per CPU – some type of CPU shielding is needed?
    - Real Time + RDMA is **very** appealing!
  - Existing softirq-thread usage design needs to be addressed
    - → with current implementation –or- do overhaul

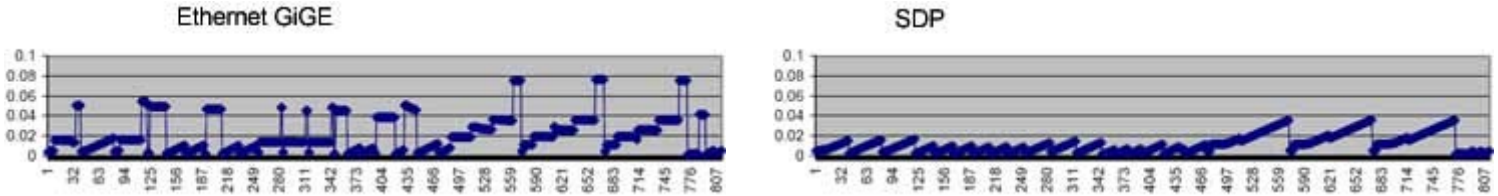
Linux Kernel Summit
Key Issues Today

- (3) Diagnostics: Linux needs more!
  - SystemTap: More Scripts, more Documentation, User Level Debugging!
  - uTRACE, PAPI
  - Concurrent's NightStar: http://www.ccur.com/isd_solutions_nightstartools.asp?o14=1
  - Tools built using Intel ATOM: http://www.intel.com/cd/software/products/asmo-na/eng/219608.htm

- (4) TCP/IP "Jitter"
  - TCP/IP & Ethernet currently demonstrate higher and unpredictable latency spikes in real life
    - With TCP/IP, we have seen latency spikes (~40ms) caused by Nagle algorithm and the congestion avoidance window
  - The jitter in packet arrival impacts decision making
  - Nagle disabled => still latency spikes
    - Ex: When connection is idle for 5 minutes (unless continuously keep setting QUICKACK which is not a viable solution)
    - SLOWSTART is unacceptable for some classes of applications
  - Could a User Space solution viable option as opposed to RDMA?

Linux Kernel Summit
Key Issues Today

- *Requirement: Need consistent low latency predictable message delivery*
  - *Ex: Socket Direct Protocol (SDP) on Infinband does not require special tuning*



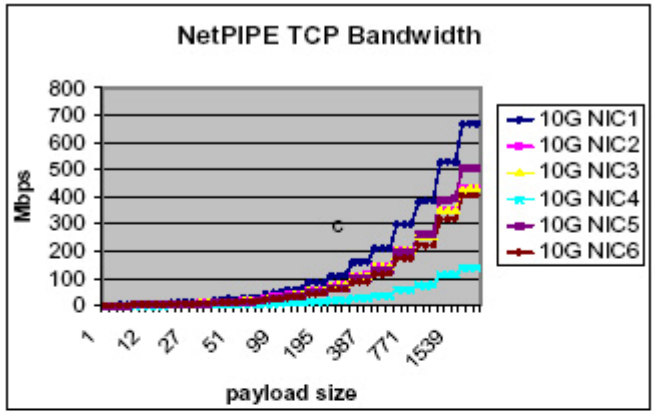
- Need predictable message behaviour, low latency, high bandwidth
  - Large-page support
    - Would help support many of the large data structures and buffers needed for supporting high-bandwidth networks
  - RDMA – seen to have benefits
    - need stable interfaces, auxiliary interfaces (ex: pin memory that is subsequently unmapped), etc…
  - TCP/IP - offloading has also seen to have benefits
    - *user space options emerging, so is it really an option? (with real time?)*
      - See: http://www.solarflare.com/technology/documents/EndoftheRoadforTCPOffload.pdf
  - Scalability of protocol processing over multi-core architectures
- Examples of some initial testing (not finished, and testing all switches- this is just one set of tests at a point in time)

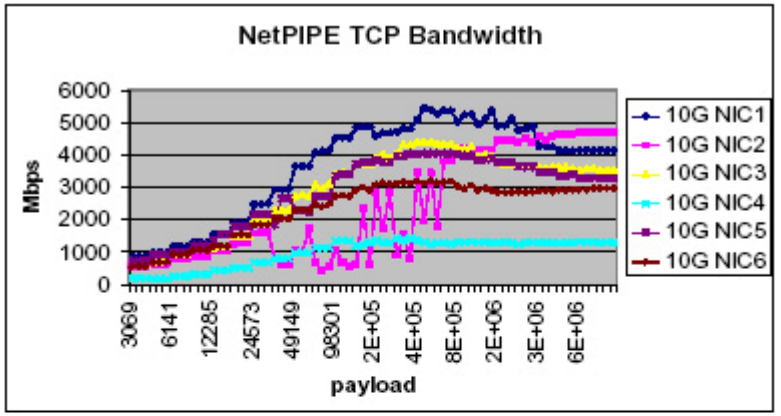
NIC Benchmark using NetPipe ping-pong mode:

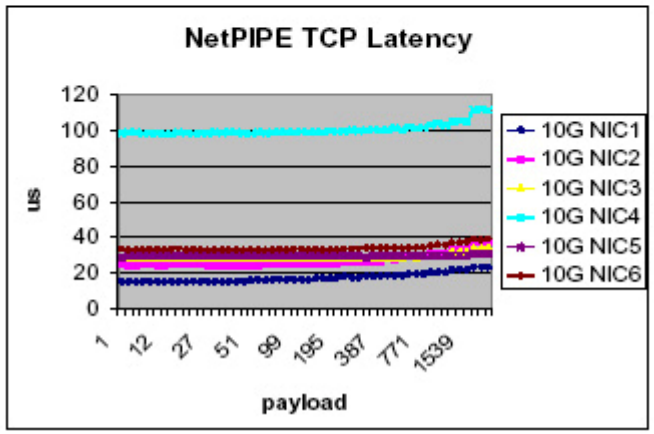
	NIC1	NIC2	NIC3	NIC4	NIC5	NIC6	iWarp1	iWarp2
64-byte NetPipe Latency	15us	24us	28us	100us	30us	32us	8us	7us
Highest NetPipe throughput	5.3Gbps	4.7Gbps	4.4Gbps	1.4Gbps	5.4Gbps	3.4Gbps	8.7Gbps	8.9Gbps

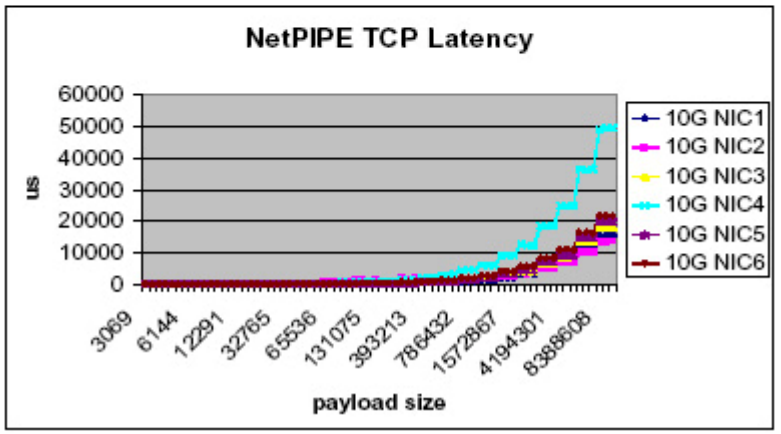
Linux Kernel Summit
Key Issues Today

– Examples continued



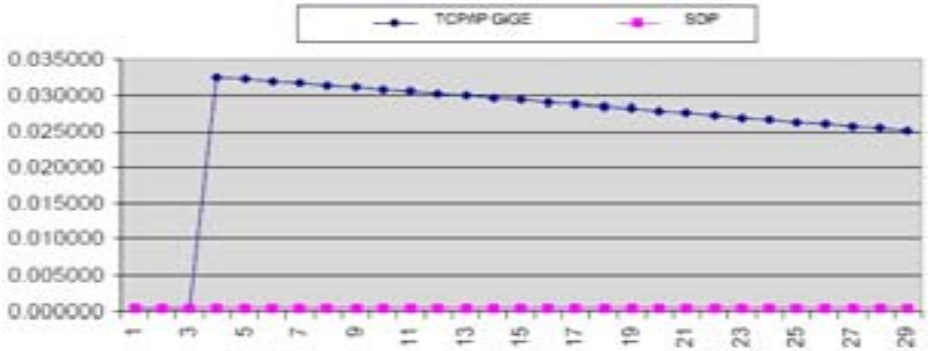






Linux Kernel Summit
Guaranteed SLAs For Bandwidth & Latency

- The desire is to have well defined, deterministic latency
  - Currently, applications send data via TCP, and we've seen high variance
  - TCP/IP Client Latency (GiGE) as compared to SDP on another fabric



- What exists today is not good enough for tomorrow – we need something better
  - The way things are done today does not make them right for tomorrow
  - Live application migration to any fabric (an option that should be considered)
  - Clock Synchronization for all hosts on a fabric
  - Ideally: Guaranteed Service Level Agreements (SLAs) for Bandwidth and Latency
  - More un-managed bandwidth is not the solution
    - wider roads invite more traffic... and ensuing congestion

Linux Kernel Summit
Why Service Oriented Fabric?

- Increase in volume of data passed between applications have created the need for lower latency solutions
  - TCP/IP has been found to be problematic for some classes of applications
  - TCP/IP was designed for wide area communication
    - Gateway window size (mtu) and "connectedness" are indeterminate

- Desire To Virtualize Entire Datacenter / Create A Service Oriented Fabric
  - Virtualization is not limited to servers
    - Server
    - Fabric
    - Storage
    - Etc
  - Migrations based on SLAs

- OpenFabrics (CS is a board member)
  - Open Standards & Open Source Solutions

Linux Kernel Summit
Q & A

- Contact Information
  - “Head” Bubba
    Credit Suisse
    IT Research & Development
    11 Madison Avenue
    New York, New York 10010
    email: head.bubba@credit-suisse.com
    head.bubba@ieee.org

CREDIT SUISSE

Linux Kernel Summit
Appendix – Testing Contacts

- See H.B. for contacts
  - Some switch/HCA/NIC vendors may make hardware available for testing with clusters using various HCAs and/or NICs

CREDIT SUISSE

Linux Kernel Summit
Key Issues Today

– Examples continued