

Workshop on Network-I/O Convergence: Experience, Lessons, Implications (NICELI)

Workshop Summary

Vinay Aggarwal
Olaf Maennel
TU München
Institut für Informatik
Boltzmannstr. 3
85748 Garching b. München
vinay@net.in.tum.de
info@olafm.de

Jeffrey Mogul
HP Labs
1501 Page Mill Road
Palo Alto, CA 94304
JeffMogul@acm.org

Allyn Romanow
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
allyn@cisco.com

ABSTRACT

This is a summary of the NICELI workshop, based on scribe reports written by Olaf Maennel and Vinay Aggarwal, and edited by Jeffrey Mogul and Allyn Romanow with help from the NICELI attendees. The workshop was held in conjunction with SIGCOMM 2003 on 27 August 2003 in Karlsruhe, Germany.

Papers and presentations from the workshop are available on the Web at <http://www.acm.org/sigs/sigcomm/sigcomm2003/workshop/niceli/>

Note-taking during an active discussion is a fallible process, so these notes may contain errors. We have encouraged participants to help us find these errors. Corrections made after this article goes to press may be found at the NICELI Web site.

1. INTRODUCTIONS

Allyn Romanow welcomed the participants to this workshop about high speed I/O and thanked the speakers for their contributions.

2. KEYNOTE: DAVID R. CHERITON, STANFORD UNIVERSITY

Network-I/O Convergence in “Too Fast” Networks: Threats and Countermeasures

The workshop began with an 1-hour invited talk by David Chertton from Stanford University. He observed that network I/O convergence is an old story. Giving a brief introduction to I/O development and its history right from the 50/60s to the present age,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

he focussed on the current problems facing the field – reordering, forged packets, replay, attack by peripherals. He expressed that the performance of the host should not be degraded by attacks.

The new problem is “too fast” networks, eg. 10 Gbps. He explained the term “too fast” to mean that “it is very expensive not to protect, and not feasible to do in software.”

He then briefly commented on zero copy – corruption and compromise, and on receiver authentication – how to do efficiently and safely.

Then, he referred to Moore's law, inferring that the too-fast networks were at the limit of their memory speeds.

He observed that there is a collision between I/O and processor for the hardware resources. There is a contention for pins, on-chip state, and on-chip logic. If the contention is not in the process, it is way across in the I/O network. He humorously stated that everybody is trying to push the other off-chip!

He went on to highlight the threat perception of Infiniband. After explaining what is Infiniband, he stated – “fix IP for storage or else lose to Infiniband.”

On the issue of multi-layer solutions, he observed that the more the layers, the more complex it will be for hardware. In the case of meta-protocols, he believed that the standards are too flexible to design hardware to a good standard. Hence, hard choices are required.

He later shifted attention to RPC, saying that since network is as fast as memory, not just RDMA, but RPC needs to be handled too.

He then proposed the solution of refactoring the transport layer protocol, based on the theory of refactoring the protocol design problem between hardware and non-hardware level.

Then he arrived on his final solution – an RDMA based protocol. He explained the term “region” as a collection of packet frames to/from which a sequence of packets of particular flow are mapped. Showing a diagram of a region structure, he talked about its delivery conditions and its pros. Tackling the issue of control, he explained the ROP control level. Then, he explained the working of File Write. For connection setup, he said that the channel manager is a mechanism to create or setup new channels.

To conclude, he reiterated the main points of his talk, and stopped with an open question – the counter-measure exists, but can the IP-Ethernet community respond?

Questions & Answers:

The first argument came from Steph Bailey: one community says TCP is good, another says TCP offload is the hardest part, not RDMA. But he believed that stable recovery and window management are the most challenging parts of TCP. What about congestion avoidance? To this, Cheriton replied that he did not address this issue in his talk. He said, perhaps, it is a good approximation.

Then, Jim Pinkerton from Microsoft said that TCP has a buffer to control that rate. The definition of the problem is interesting. He disagreed with Cheriton's statements about Infiniband. In Infiniband, the layers work differently. While the transport layer is very similar, the reliable layer is quite different. The problem is about congestion control and new algorithms.

Cheriton responded, all that has been learned about/from TCP can be applied again. All the mechanisms and problems about TCP can be solved by designing a new protocol. However, one needs to keep an eye on cost. The implementation of stacks on hardware needs to comply with cost constraints: it needs to fit on the chip.

Then Renato Recio made 2 comments: the discussion is specific to 1 protocol, the problems are not at the layer that the speaker is talking about, marshalling-demarshalling happens in Java, C#, etc. The answer was that RPC parameters must tie into RDMA for success. The dynamics here look similar to TCP. Due to the amount of memory wasted, the same ideas would have been comical 20 years ago, but not anymore, as memory is cheap now.

Jeff Mogul, HP Labs made the next point. He said that the frames are based on MTU size. If MTU changes, configuration changes will be required which the attacker can exploit. Cheriton agreed with this.

Next, there was an extended interaction between Jeff Chase of Duke University, Jeff Mogul, and David Cheriton. Jeff Chase began by observing that there is a collision between research and industry view. The research community believes everything can be built from the ground-up, but industry does not like this. It wants a standardized IETF blessed protocol. They would prefer to build on top of an older protocol. To this, Cheriton interjected that betting on TCP is bordering on the absurd. Then Jeff Mogul chipped in saying that problem lies in integration with the OS, not with TCP. Cheriton said that demands of what we build are colliding with the hardware reality. He then discussed TCP and SCTP. He remarked that the tipping point has arrived, and hence, change is essential.

Brent Callaghan from Sun Microsystems referred to the simplified RPC protocol. When RPC runs over any transport protocol, it is not so good. However, running over RDMA is quite a refreshing change.

Dave followed this up by saying that much of the proposal is for file access. When RDMA token is passed to the server, and server turns it back. This is an alternative way of flow control, and does not commit so much memory. He insisted that this is only a comment.

Cheriton concluded the questions by announcing that one may add to something, but, it is more complicated and expensive.

3. SESSION 1: PROMISES AND REALITY

3.1 Renato Recio, IBM

Server I/O Networks – Past, Present and Future

The speaker is associated with Infiniband, RDMA and IBM eServers at IBM Research. Recio said that the agenda of his talk was server I/O – network types, requirements, I/O attachments, etc. After highlighting the purpose of server I/O networks, he talked

briefly about server I/O network requirements – standardization, performance, high availability, low cost – to name a few. Then he added to this list – virtualization, security, and service differentiation. Then he went on to summarize the server I/O network history and the network evolution timeline.

The speaker then switched to PCI. The strategy – add evolutionary technology enhancements to the standard. He maintained that there are 2 contenders: PCI-x and PCI-express.

After comparing the various I/O attachments, he spoke at length about Infiniband (IB) model and its strategy. When he compared IB with PCI-express, Raj Yavatkar objected saying that the comparison was inappropriate, as both have different uses.

The next part of his talk was on server scale-up technology options, server IOA outlook, attachment and expansion. Mentioning the problems with sockets over TCP/IP, he outlined the basic mechanisms of network offload. After explaining in detail the IB network stack offload, he showed the same for iONICs.

He then extolled the network offload benefits from the middle-ware's view. He went to say, rather artistically, "TCP/IP/Ethernet are kings of LANs!" Then, he devoted some attention to LAN issues, before discussing cluster network contenders, proprietary and standard networks. In the end, he briefly gave an overview on HPC cluster network outlook.

To summarize, he observed that I/O server adapters will likely attach through PCI family. He then gave a brief outlook on what each kind of network will likely use and why.

Questions & Answers:

Joerg Micheel (Endace) expressed interest in latency requirements, and desired to know where these came from. The reply was that there is no paging and congestion to hold that off.

David Cheriton was surprised that latency was pushing down through switches. He was not convinced that locking, interaction, etc. should have that kind of overhead. He believed that the logic for latency and overhead was not good, as disks and other old-fashioned devices were being used for paging, which were not appropriate. Recio contended that each device in the path had latency, and that lowers the number of operations. When questioned why, Recio referred to utilization of path, throughput and latency function. He reflected that lower latency implies less hardware and lesser cost.

Donald Newell from Intel remarked that this is a data-free argument. There are ways to hid latency with software. He was looking for a concrete data on this argument. He would like, for demo purposes, a set of applications for latency-sensitive productivity. Recio said, he would provide the required data.

3.2 Piyush Shivam, Duke University

On the Elusive Benefits of Protocol Offload

Co-author: Jeff Chase

Piyush began by introducing the offload controversy and NIC cards. He outlined recent technology trends and Moore's law, with an illustrative graph. Some minor doubts were raised on the correctness of the graph, which Jonathan Smith strengthened by declaring that the graph data is incorrect. Piyush offered an explanation, which was accepted!

The speaker continued with application trends, and prepared the ground for LAWS model. When he presented the LAWS ratios, Renato Recio objected that 1 ratio was missing, that for offload. Piyush said it will come later.

Piyush observed that LAWS captures application trends. Analyzing LAWS, he pushed for ignoring latency, and laying stress on

throughput speed. He went further to present some algebra to press his point.

When he pointed out the benefits of host-limited case, Raj Yavatkar wondered if it really mattered for network-intensive cases. The answer was deferred till forthcoming slides.

He then explained the benefits of network-limited cases with a graph. His cardinal argument was – for very fast networks, this case have very good benefits, but they are valid for a very few applications only.

He went on to counter the question – will network bandwidth outrun Moore's law?

When he approached the NIC-limited case, Renato Recio pointed out that some NIC designs can do what high-end systems do. Recio's question was if this particular case was covered by their suggestion/theory (of lag ratio). Piyush replied saying, one has to be limited either by host, network or NIC in any case, and anyway, that is covered.

After giving an overall picture, Piyush concluded that applications need to be understood to understand the role of TCP, IP of-flood, RDMA, etc. He also believed that point studies are misleading. He stopped after giving a brief LAWS analysis.

Questions & Answers:

Joerg Micheel from Endace started off by announcing that he is designing high-speed network interface for PCs. He operates on the parameter that the capability of CPU to process packets is the main factor. He then stated that the presented model is totally incorrect, as the factors are wrong. He believed that it does not matter what processor is present, the limiting problem is the bandwidth of the machine. Hence, the model does not fit. Piyush replied that he has considered the end-to-end throughput, and hence Jeorg's concerns are covered. The parameters do suffice. Joerg again objected that it has nothing to do with CPU speed, but with I/O problems! At this point, Jeff Chase intervened and said that the model is a simplification. the problem manifests itself in CPU speed. There exists no assertion that, by doubling the CPU speed, one can cut the latency. Hence, the model works. Steph Bailey seconded the arguments in favour of the presentation throughout.

3.3 Samuel Fineberg, HP NonStop Labs

Performance Measurements of a User-Space DAFS Server with a Database Workload

Co-author: Don Wilson

Sam began by explaining DAFS. After giving the characteristics of direct access transport, he gave some DAFS details. He compared inline I/O with direct I/O. He then briefly went over Oracle disk manager, prototype client/server, test system configuration, his experiments, ODM blast, its read and write comparison, ODM latency test and performance.

The speaker then gave some Oracle-based results and the Oracle TPC-H performance. He also pointed out that in the operation distribution, read operation was the largest component.

He later concluded that local I/O is still faster, that DAFS still has more capabilities than local I/O, and that memory registration is yet a problem with DAT.

Questions & Answers:

Jim Pinkerton queried if the speaker had a feel for memory registration bottleneck. Sam replied that the implementation was in hardware. When Jim offered that good caching algorithms exist, the speaker expressed his reservation saying, the problem with caching is that VM tables change often. OS cooperation is also required.

4. INVITED TALK: WU-CHUN FENG, LOS ALAMOS NATIONAL LABORATORY AND OHIO STATE UNIVERSITY

Bridging the Disconnect between the Network and Large-Scale Scientific Applications

The speaker started off by introducing the Green Destiny super-computer. He presented the grand end-scheme: you put a CD (containing the software you want you install in the whole network) into your cluster/network, the software figures out everything automatically, and installs a clustered software in your network. Then he asked the question: does our network possess similar virtues?

Then he made a very humorous reference to the issue of how many staff was required to get a simple Internet connection started at SIGCOMM 2003 conference. He enumerated the various system failures and correlated them to non-transparency in Internet, using terms like NAT, DNS, etc. His main point was, even getting a simple Internet connection running involved knowing so much technical terminology, that it was very difficult for a layman to get it done on his own. The ironical comparison of the problem with the conference Internet connection was well-taken by the audience!

He then elaborated the argument that complete transparency is missing in the network. He observed: "Why can't I just plug my cable into the wall and get my Internet connection running?"

He explained what he meant by "disconnect."

At this point, Renato Recio stated that the requirements of Lawrence Livermore are much more complex than the speaker's. He asked if the speaker had investigated that case. Wu replied that he had not considered that data with respect to the nuclear program. He further stated his belief that in all applications that he came across, the latency issue can be cushioned if it is not ... To this, Recio countered that latency interacts with throughput. As Feng was about to reply, Jeff Mogul intervened and terminated the discussion due to time constraints.

Feng went on to explain the wizard-gap problem. He concluded his solution by another humorous comparison of its performance with that of FedEx, and emphatically stated that his solution beats FedEx by speed and throughput, even though it requires around 30,000 professionals working at 4 different sites.

He finished his talk by briefly talking about the dynamic right-sizing.

5. SESSION 2: STORAGE PROTOCOL DESIGNS

5.1 Brent Callaghan, Sun Microsystems

NFS over RDMA

Co-authors: Theresa Lingutla-Raj, Alex Chiu, Peter Staubach, Omer Asad

The speaker commenced by explaining the need for RDMA as a transport layer protocol. He commented that NFS is an RDMA sweet-spot. He promoted RDMA as a new RPC transport. After briefly explaining some small RPC messages, he explained the moving of NFS data with RDMA. He later compared NFS throughput with that of TCP and RDMA. He concluded his talk by explaining the extended RDMA transport header.

Questions & Answers:

David Cheriton reflected that the speaker started with read-read protocol, while he himself had started with write-write protocol.

Combining the two, we now have a read-write protocol. He personally hated RTTs.

Brent replied that he too had started with write-write, but found it very complicated, he experienced problems with the server. Hence, he moved over to read-read. The main problem was that priority buffers could not be figured out. While read-read seemed much simpler, write-write is definitely more practical.

Jim Pinkerton asked why NFS/TCP performance appeared better than NFS/RDMA performance for small transfers. Brent explained that RDMA memory registration impact is greater for small transfers. A new read-write protocol will allow the client to reduce memory registration overhead.

5.2 Mallikarjun Chadalapaka

A Study of iSCSI Extensions for RDMA (iSER)

Co-authors: Uri Elzur (Broadcom), Michael Ko (IBM Almaden Research Center), Hemal Shah (Intel), Patricia Thaler (Agilent Technologies)

Questions & Answers:

Austin Donnelly from Microsoft queried if the proposal does not open an attack.

Mallikarjun replied that there are other iWarp mechanisms to deal with it. One example would be invalidating the S-tags. The damage is localized, some other I/O will be aborted, but it stops at that.

At this point, Jim Pinkerton, the Session Chair, intervened saying that there is a detailed analysis of security for RDMA in an IETF Internet-Draft.

David Cheriton raised the next question. The copy overhead does not register, there is a long history of dealing with virtual memory systems. We do not force revalidation, and some other such things. Then why not use the VM model?

The speaker replied that they do not have command parts, only data movement is involved. When response comes back, buffer associated with transaction will not be minimum. VM can be used locally.

Pinkerton said, should we incorporate iSER into VM?

Cheriton asked, why not use the same VM? There are a lot of synergies, and common software and techniques.

Pinkerton offered, it is functionally equivalent to VM, but not synchronized with local system. That is the main problem.

Ted Kim changed the topic by asking if the speaker had considered using IB. The speaker replied that iSCSI is defined over TCP. He wasn't sure if iSCSI can run over IB. He explained his position by stating his design goal – to allow iSCSI to run on generic RNICs specifically. One can always do vendor-add ops. Efficiency on generic cards is important.

Renato then commented that integration of this chunk with the processor may allow for pre-activity: processor with paging or virtual and physical mappings.

Pinkerton concluded, we saw new innovations in IB. We observed that the scope of the protocol expanded. The new development is that RDMA over TCP runs in the kernel in an optimized way. It would be interesting to note how more applications map with RDMA.

6. SESSION 3: NOVEL APPROACHES

6.1 Angelos Keromytis, Columbia University High-Speed I/O: The OS as a Signalling Mechanism

Co-author: Matthew Burnside

Before commencing his talk, the speaker elucidated that he hails from a cryptography background, hence his paper/talk constantly refers to cryptography, and is based for the most part on it.

Questions & Answers:

Jim Pinkerton offered that memory is a concern because it is a peripheral. If we resort to using a buffer, the problem will dissolve. This suggestion was backed by many other people in the audience. The speaker admitted it partly.

Donald Newell from Intel said, complete parallelism has been achieved in the proposed method by moving memory to other devices, and then employing additional complexity to manage it. A better solution would be to employ memory parallelism closer to the process. More random external accesses will be available, and better performance will result. There are better methods available, more scalable and cheaper.

The speaker replied that he looked at programmable FPGA's, they weren't that complicated. However, he confessed that he didn't have the figures for performance comparison.

Newell then asked, what kind of applications the speaker would be interested in demonstrating his work – databases or simpler applications?

The speaker countered that Newell was thinking of too high a level of complexity. He does not want any SQL queries, but a much simpler process.

Renato then commented that other control operations would need to be appended. The speaker would need some kind of specialized control, etc. thereby making the whole scheme very complex. In other words, he would not be able to escape a certain degree of complexity anyway.

David Cheriton said, what the speaker described for peer-peer is all transforming into a network, and then, the speaker walked into another discussion. But Cheriton questioned if it runs over Ethernet, for he believed, in the end, its all just Ethernet and some devices.

The speaker agreed that the next step indeed would be to see Ethernet, as there were many interesting issues involved there. However, he maintained that all the devices are interconnected, and one may run IP over it.

6.2 Kieran Mansley, University of Cambridge Engineering a User-Level TCP for the CLAN Network

The speaker started off by explaining his problem – the networks have become faster, and so have the transmission speeds of packets, but the overhead to deal with the packets is still the same. And this requires CPU cycles. How do we tackle this?

After referring to some possible solutions, he arrived at CLAN networks, and explained in depth its user-level stack architecture.

Jim Pinkerton asked for some figures regarding bandwidth and latency, which the speaker promptly supplied.

Then the speaker spoke on true zero-copy transmission.

He later summarized his talk saying that the TCP/IP stack should be moved to the user level, and that the retransmission should be handled by the gateway.

Questions & Answers:

Renato Recio argued that the speaker might still need a copy operation to remove headers from the stream, only that now it would be middleware headers rather than TCP/IP headers. The speaker agreed.

Jim Pinkerton first commented that only the transmit path has been considered. Then he asked what the speaker meant by an “asynchronous API.” The speaker replied that the protocol processing will have to be done at some point, either immediately at the time when asked, or sometime later, even if the application is blocked now, or later.

Stephen Bailey doubted if the socket API really required the copying. He expressed that the speaker had not changed the socket API per se, but only the interface (in a rough sense), according to people’s expectations. He summarily believed that what the speaker had done was not really required. The speaker replied that the kernel could spend a lot of time servicing the queue, and as a result would have to perform a copy without these changes.

6.3 Rolf Neugebauer, Intel Research Cambridge

A Case for Virtual Channel Processors

Co-author: Derek McAuley

Questions & Answers:

Mallikarjun Chadalapaka raised the first question. He surmised that constructing the VCP would involve a tremendous amount of effort – to take the TCP stack and iSCSI and then make the VCP. He asked if the speaker could specify what work was involved in defining a VCP.

The speaker agreed and said that he was currently in the process of doing it. He confirmed that it did involve a TCP stack, a device driver, and putting it together with an API. However, he believed that a minimal OS running on top of a VCP is enough to get a TCP implementation running.

Chadalapaka further queried if in the case of iSCSI resources, one needed separate NICs for TCP and SCSI. The speaker replied that there exist NICs that multiplex on a large level. Many new NICs have been introduced of late. They later demux to different functions.

Jeff Chase then reflected that the first talk proposed to add a bunch of data copies, that would certainly require a lot of support in VM. The speaker agreed, but said that the VM already does it, it even has an interface for this function. When Chase asked if he had full control of VM between paging, he replied in the positive.

7. GENERAL DISCUSSION

All attendees were requested to air their views on appropriate issues.

Jonathan Smith said, are there any economic factors to change the bias in memory architecture towards faster memory rather than larger memory? He further said, do we always have to work around it, or can we hope for something better.

Stephen Bailey proposed stream benchmarks. He said it is faster, exponential. The number of pins is a limitation, it does not grow. And so is bandwidth. But the stream is quite good, and fast. We have 20 years of stock. Memory will be the bottleneck.

David Cheriton expressed his inability to comment on physics. However, he said he would comment on old perspectives. He would like to get rid of that copy overhead. It should be easy to transition to new, faster mechanisms. We do not notice success due to page manipulation. He has a vision to see 10 Gig Ethernet coming directly into the processor, and memory systems being accessed through VM, paging, and so on. There is not a drag behind, but it gets eaten up by little things like checking this or that. To fold in security, the complexity must be reduced, so much so that it is

understandable by my mother. We still are a far way from that.

Jim Pinkerton would like to see network speed approximate memory speed. He believed that enough technology is lying around for 20 years. The network is no longer the orphan child, it gets the first technology. It has the same technology as does a CPU now.

Steph Bailey commented that disk channels have consistently delivered 10x the bandwidth of NICs to the same system. With the right hardware, NICs should be able to achieve parity with disk adapters (picking up a factor of 10 performance) without any host hardware changes.

David Cheriton said that in the network world, people talk about latency sensitive applications. This market is worth \$800 billion, comprising block storage I/O, fibre channel, whatever. Data management needs to be centralized. How does one run Oracle on a Sun server 1000 miles away, when the fibre is cheap and has a capacity of 40 Gig? How do we design a protocol that works reliably, securely and does not get overloaded. That is the grand challenge, that we have to do right. After all, there will always be 1 orphan child.

Jeff Chase doubted if it really is a latency-driven world.

Cheriton reconfirmed that it is, down to the level of 100 microseconds range. When dealing with disk I/O, the numbers are in milliseconds range.

Jeff Mogul suggested that if we replace disks with MEMS devices, then we might end up on a better price curve. In that scenario, 1 ms will be important, and maybe even 100 microseconds.

Jeff Chase suggested that most systems are throughput-sensitive, not on latency. For IOP, Internet requests served per second is the sensitivity factor.

Jim Pinkerton remarked that everything is opinion-based, and not fact-based. He raised the question that what areas need more research.

Donald Newell backed him, saying that this is the most common argument. People often come up with bizarre requirements.

Renato Recio commented on the areas for additional research. According to him, segmenting out the problem is oversimplification, and that is the problem. There is a big problem with Java, C#, etc., how do we get advantage in these areas. It addresses a large chunk of space, and involves a lot of money. It might even take application changes.

Jim Pinkerton expressed that whenever application change is required, it introduces a new world. No features exist for it.

Jeff Chase referred to the talk by Piyush Shivam, and asserted it is all facts, not mere opinions. Once we can figure out the values for the parameters, we can visualize real environments. The problem is that we do not know what really matters, and hence estimating the parameter values is difficult.

Stephen Bailey cautioned all to be careful to go out and test something.

Jeff Chase spoke about some service benchmarks, saying that he did some work there.

Jeff Mogul then said that we all have focussed a lot on performance. The theory is that if we employ RDMA, we get better performance. But he believed that the right way to look at it was, does RDMA enable systems vendors to get acceptable performance with cheaper hardware? If noone uses hardware-RDMA adapters, is RDMA really better? What is the right question to ask – faster, faster... or cheaper, cheaper...? He firmly believed that cheaper is better than faster.

Jonathan Smith first expressed agreement for the previous comment. Then he said that our research is greatly hindered by absence of data, especially in fields of security. He urged all if we could come up with an anonymization policy to facilitate better availab-

ility of data.

Brent Callaghan drew attention to the fact that use of sockets was primarily responsible for the popularity of Ethernet. He then asked if we have the right API for RDMA as yet. It may be RPC, or it may be MPI. We do not know for sure.

Renato Recio then said that it takes time for applications to develop.

Jeff Chase commented that in the short term, it is more economical to buy more servers rather than hire people.

Lastly, Kieran asked how to make the API better for applications.