印度 Bangalore 之 HPCA/PPoPP-2010 与会小记

包云岗 2010-2-1

中科院计算所系统结构重点实验室

一、会议简介

HPCA（IEEE International Symposium on High-Performance Computer Architecture）是计算机体系结构领域的最重要会议之一，创办于 1995 年，由 IEEE 组织。经过十多年的发展，目前已经和 ISCA、Micro 一起誉为体系结构领域三大会。HPCA 的截稿期在每年的 8 月份，一般录用 30 篇文章，录用率在 15~20%。2010 年的 HPCA 在印度 Bangalore 召开，一共录用了 32 篇文章，录用率为 18%。

PPoPP （ ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming）是并行计算领域的最重要会议之一，创办于 1988 年，由 ACM SIGPLAN 组织。并行计算领域很多经典的工作都是首先在 PPoPP 上发表的，例如 BSP 模型、LogP 模型等。随着 Multicore 技术的兴起，并行计算也越来越重要，因此从 2005 年起 PPoPP 由每两年召开一次改为每年召开一次。

2008 年，HPCA 与 PPoPP 采用了联合举行的模式，使这两个会更像是一个 Multi-Track 会议——两者的 Keynote 和 Panel 是在一个会场，而之后的会议报告则是在独立的会场。这种模式取得了非常好的效果，更利于体系结构研究人员与并行计算研究人员对彼此领域研究方向的了解和交流。所以，我还是很喜欢这种多个领域共同举行会议的方式。大家还可以关注每年的 ASPLOS（体系结构、程序语言、操作系统交叉的会议），每四年一次的 FCRC（ISCA、PLDI、STOC、Sigmetrics 等 ACM 众多 6 月份的会议联合举行，下一届是 2011 年 。

图 1. HPCA/PPoPP 会场

我这次和清华大学计算机系的翟季冬一起前往 Bangalore 参加会议。翟季冬是陈文光教授的博士生，今年他们有一篇 PPoPP 文章。怀着对会议交流的期待和对印度风情的好奇，我们 1 月 9 日清晨从北京出发，由新加坡转机，于 Bangalore 当地时间晚上 11 点左右抵达酒店安顿下来。

二、Workshop

1 月 10 日用完早餐，我们在等待大巴的时候碰到 UIUC 的 Wen-Mei Hwu 教授，我上前

用英语和他打招呼，介绍自己是来自中科院计算所，没想到他跟我说“我们可以用中文聊”，一下子感到特别亲切。在大巴上我和他坐在一起，我们聊了很多，比如计算所情况、龙芯大公司、曙光 6000、印度印象等。他也提到这两天正在参加 Indo-US Workshop on Parallelism and the Future of High-Performance Computing。抵达会场，我也参加了这个 Workshop。

2.1 计算机教育

上午的第一个话题是计算机教育，嘉宾都是印度学者与美国学者。给我印象非常深刻的是一些数据：印度大约有 15 万学习计算机专业的学生；但整个印度只有 15 个好的大学才能培养博士；印度现在大约有 750 名在读博士；印度每年大约有 100 名计算机专业的博士毕业！但与之形成鲜明对比的是有大量的印度学生在美国攻读博士学位。印度学者认为，缺少好的师资是最大的问题，并希望能与美国加强合作，联合培养更多的学生。

美国学者，如 UT Austing 的 Yale Patt 教授等，则是对现在的计算机教学现状提出了质疑，认为很多课程并没有传授一些最基本的思想与概念，而仅仅是一些具体与复杂的例子。以编程语言为例，很多大学都只是教授一门具体的语言，C/C++/Java，却不开设程序设计语言原理这样比较基础的课程。在讨论的时候，有嘉宾甚至拿 MIT 的 Frans Kaashoek 教授开设的 6.033 Computer System Engineering 课题开刀，认为这门课程只是在教给学生一些具体而复杂的例子，还要做大量的项目，给学生带来巨大的负担，却没有传授一些基本的原理。这个说法立刻招来一位现场 MIT 教授的反驳，他认为 6.033 这门课在所有 MIT 计算机课程中都是一门非常棒的课，它体现了理论与实践的结合。

我个人还是很认同 MIT 那位教授的观点，计算机系统结构领域具有很强的实践性，尤其是要在这个领域开展研究，很多时候都是“实践先行、理论跟上”。而 Frans Kaashoek 开设的 6.033 这门课，实际上很好地体现了系统结构这个学科的这种特色。当然，我也很赞成计算机本科课程要多给学生灌输些各个领域的基本概念，以扩大知识面。

2.2 工业界研究动态

第二个主题是关于工业界的研究，嘉宾是 HP、IBM、Intel、AMD 和印度 TATA 公司的一些专家，他们分别介绍各自公司正在开展的或者关注的一些研究领域。

图 2. HP 的 Intelligent Infrastructure

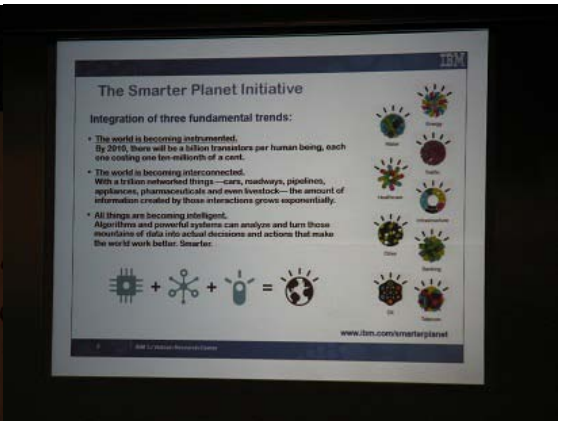

图 3. IBM 的 Smarter Planet Initiative

例如，（1）HP 提出了要构建 Intelligent Infrastructure，目标是研究一系列新颖而激进的数据收集、存储、分析和传输技术，以满足 Exascale 数据中心对数据的需求。他们的研究主要关注可管理（Manageability）、高效（Efficiency）和可编程（Programmability）。（2）IBM 认为，不管是科学数值计算还是企业应用，传统的大规模高端计算（High-end/large-scale）

都是和高性能计算（High-Performance Computing）密切相关的。IBM 判断最新的趋势是人类的日常活动也将和高性能计算融合，因此他们提出了一个雄心勃勃的 Smarter Planet Initiative（智慧地球计划）。（3）Intel 主要介绍他们在云计算方面的研究，包括前段时间刚发布的 48 core 的芯片、面向多核编程集成工具的 Parallel Studio 等。这些企业都非常重视与大学、研究机构的合作，每个企业都有自己的高校合作计划。

Wen-Mei Hwu教授提出了一个合作的知识产权问题，引起大家激烈的讨论。观点基本分为两派，企业代表认为知识产权不是大问题，比如他们也支持一些开源项目，但有些大学教授还是认为知识产权问题还是会影响合作。总体感觉，这个Indo-US Workshop其实就是一组Panel，嘉宾基本上包括所有参加HPCA与PPoPP的大牛们，比如Stanford的kunle Olukotun、UIUC的Wen-Mei Hwu、MIT的Arvind、Univ. of Michigan的Trevor Mudge、Wisconsin的Guri Sohi、UT Austin 的 Yale Patt 等 。（其他信息可访问： http://polaris.cs.uiuc.edu/ppopp10/indo-us-workshop-program.html）

三、Keynote 与 Panel

3.1 重量级嘉宾——印度前总统

10 日晚上 5 点半接待晚会才开始，我们下午离开会场，在附近晃了一圈。当 5 点半多返回到会场时，发现多了不少持枪的军人。当我们走进报告厅，才知道原来印度前总统Dr. A. P. J. Abdul Kalam（阿卜杜勒.卡拉姆）也来了，并会作一个报告。很遗憾，由于Kalam的浓重的印度口音，我听得云里雾里。好在他的主页上有这次报告的演讲稿[1]，我才在会场之后了解到他演讲的具体内容：

印度发展高性能计算机也是因为发达国家的禁运。印度是从上世纪 80 年代才开始研制并行计算机，于 1986 年有国家航空实验室（National Aeronautics Laboratory, NAL）研制成功第一台并行计算机 Flosolver MK1，最新一代是 MK6，由 128 个 Pentium 构成。印度研制的其他并行计算机的机构主要有 ANURAG (Advance Numerical Research and Analysis Group)、BARC (Bhabha Atomic Research Centre)、C-DAC (Centre for Development of Advanced Computing)。目前印度最快的计算机室 2007 年 10 月由 TATA SONS 的计算研究实验室研制的 EKA 超级计算机，号称当时世界 Top500 排第八，亚洲最快。EKA 的峰值性能为 172.2TFlops，实测 Linpack 持续性能为 132.8TFlops。Kalam 对未来高性能计算机体系结构展望主要突出要利用纳米技术、降低能耗和碳足迹（Carbon Footprints）、减少电子垃圾实现绿色计算等。

Kalam 在印度也是一个传奇人物，是印度最具名望的科学家之一，负责研制印度第一枚火箭、第一枚导弹、第一颗核弹等，被誉为“印度导弹之父”。之后他又步入政界，并于 2002 年当选为印度总统。（想了解 Kalam 更多信息可以访问他的主页：http://www.abdulkalam.com/kalam/index.jsp）

[1] “High performance computing and challenges”, address during the inauguration of Symposium on High Performance Computer Architecture &15th ACM SIGPLAN Symposium on Principles – Practice of Parallel Program at IISc Bangalore. Abdul Kalam, India President.
http://www.abdulkalam.com/kalam/jsp/display_content.jsp?menuid=28&menuname=Speeches%20/%20Lectures&linkid=68&linkname=Recent&content=1445&columnno=0&starts=0&menu_image=-

图 4. 印度前总统 Kalam 作报告

3.2 Exascale Computing

11 日上午大会正式开始，第一个 Keynote 是 IBM 的 Tilak Agerwala 作的报告“Exascale Computing: The challenges and opportunities in the next decade”。报告首先从分析几个潜在应用出发，说明 Exascale Computing 的必要性和重要性，接着提出了几个挑战。但这些挑战无外乎功耗、编程、可靠性这些众所周知的难题。这个 Keynote 整体上并没有给我留下特别深的印象，但其中一些细节我觉得有一些意义：（1）Exascale 将用于一百万个节点，按照现在的数据计算得到的平均无故障时间（MTBF，Mean Time Between Fault）只有 3 分钟！（2）他认为利用 checkpoint 来提高可靠性非常困难，因为 checkpoint 自身的开销是非常大（比如消耗内存带宽），所以应用也需要支持可靠性，但并没有提到基于算法的容错技术（ABFT，Algorithm-Based Fault Tolerant）；（3）Exascale 的软件依然是一个非常大的挑战，一些传统的优化技术需要改变，例如编译优化技术应该以能耗为新指标。

11 日下午有一个 UIUC 的 Josep Torrellas 主持的 Panel——Extreme Scale Computing: Challenges and Opportunities，嘉宾包括 UIUC 的 Bill Gropp、Rice 的 Vivek Sarkar、IBM Watson 实验室的 Jaime Moreno 和 Stanford 的 Kunle Olukotun。总体来上，这些嘉宾都把 Extreme Scale 认为是 Exascale，所提出的挑战与机遇和上午的 Keynote 基本一致，当然在某些方面有所侧重。有一些观点是好几位嘉宾都提到的，例如，异构硬件、专用执行单元是提高性能功耗比的最有效手段之一，存储结构应暴露给程序员以充分利用应用局部性，领域专用语言和领域知识能使开发人员具有更高的开发效率、使系统具有更高的执行效率等。

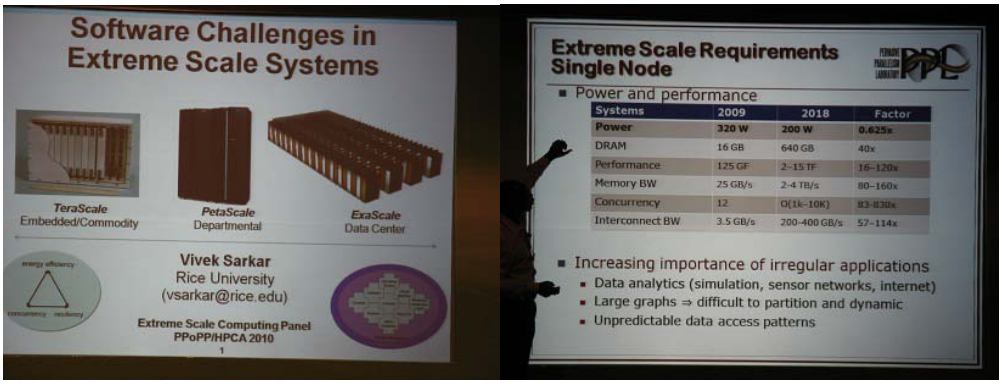

图 5. Extreme Scale Systems

3.3 Is Hardware Innovation Over?

12 日上午的 Keynote 是 MIT 的 Arvind 教授，题目是“Is Hardware Innovation Over?”。

Arvind 教授认为，当前计算机领域的研究正在从 PC 转向手机，而这和 80 年度初从大型机转向 PC 的情形非常类似。他认为，未来计算将主要由两部分组成，一部分是便宜但功能强大的前端手持设备，另一部分是为手持设备提供服务的后端设施。对于前端设备而言，当前的芯片尽管能满足高性能、低功耗，但是设计与实现都很复杂。传统观念认为芯片设计本来就是一件很困难、高风险、不灵活、易过时、易出错的事情。Arvind 认为今天的硬件设计就像是上世纪 50 年代高级语言还没发明时的编程方式，那时程序员需要了解几乎计算机的所有细节才能编程。但如今，已经有一些高级硬件综合工具可以掩盖很多硬件设计的细节了，从而降低硬件设计的复杂度。

他认为，新的并行软件目标应该是使用最少的资源来达到所需要的性能，而不是产生上百上千的线程以尽量让所有的处理器都忙碌。所以，比较高效的方式就是 Multicore 的每个核可以用来实现某种特定功能；而“如何为这些核编写应用程序让它们能实现特定功能”就类似于硬件设计（定制这些核）。Arvind 提出他们研究的 Bluespec 正是为了解决“如何高效地定制核”这个问题，而且比当前的 SystemC 等方案更有效。他展示了几个利用 Bluespec 快速实现定制芯片的例子，例如，在 24 小时内利用 TSMC 的.018 工艺完成一个 802.11a 无线传输芯片设计；只用 8 千行 Bluespec 代码、8 人月就实现了一个 H.264 视频解码芯片，支持 720p@32FPS。

至于这个 Keynote 题目所提出的那个问题的答案，既然这个问题是由 Arvind 自己提出的，那么答案自然也很明了——Hardware innovation is far from over！

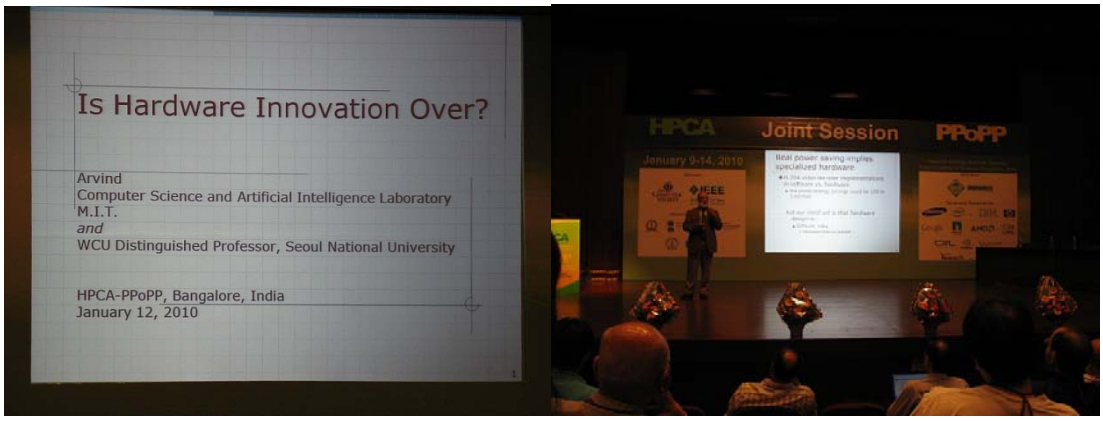

图 6. Arvind 的 Keynote

3.4 关于 Keynote 与 Panel 的困惑

除了 Arvind 的 Keynote，其他几个 Keynote 与 Panel 并没有给我一种印象深刻的感觉。我自己也很困惑，Keynote 与 Panel 都是大牛们在高瞻远瞩，是一个会议最精彩的亮点，为什么我却没有感觉呢？我觉得自身可能的原因至少有两个，一是相关的背景知识还是不够，特别是在程序语言与并行化方面，二是听力还有待提高，特别是要适应印度人的口音。

这次会议的 Keynote 与 Panel 几乎都没有提及 Exascale Computing 与 Multicore 对 OS 设计带来的挑战与机遇。是因为这些大牛们觉得 OS 不那么重要，还是因为他们都不是 OS 专家所以不关心 OS 呢？

四、Best Paper Session

大会第一天（11 日）Keynote 结束后紧接着就是 HPCA 的 Best Paper Session。

HPCA 的 Best Paper 评选非常规范。早在开会前一周，Program Chair 就把 Review 得分最高的 4 篇文章发给所有注册会议的人，提醒大家务必领取选票、参加 Best Paper Session 并在 Session 结束后投票。现场报道时，组织者除了给我们论文集、礼品等，还给我们一张卡片，上面有 4 篇文章题目与作者，这就是选票。Best Paper Session 是由 UT Austin 的 Yale Patt 教授主持的。他规定每篇文章报告时间不能超过 23 分钟，但观众提问没有数量和时间限制，只有你觉得这个问题对投票有帮助，就可以问。

第一篇文章是由来自 UT Austin 的一位中国人 Lei Chen 报告，他们的工作主要是优化间接跳转预测，思想简单而巧妙，报告也是一气呵成，非常精彩。接下来的问题却是非常的尖锐，有人问优化效果的评估，有人问实现细节。我还记得 Yale Patt 问了问题，他质疑这个工作为什么没有和最好的分支预测技术对比。我觉得这似乎是一个很难回答的问题，不过好在 Lei Chen 早做好了这个问题的准备，翻到后面的 Backup Slides，引用 06 年一篇 ASPLOS 文章的数据做了很合理的解释。

第二篇文章由来自 Intel 的另一位中国人 Tong Li 报告，是研究 Overlapping-ISA 异构平台上的 OS 设计。报告的思路很清晰，首先分析 Overlapping-ISA 是 Performance Heterogeneous 和 Functionality Heterogeneous 的结合，而 OS 支持这种异构有很大的挑战，进一步提出了一系列解决这些挑战的方法，比如通过线程迁移解决正确性、通过调度解决性能异构问题。他们在一个真实系统上构造出这种异构模式，并在 Linux 2.6.24 内核中实现了他们的方法进行评估，工作非常有说服力。报告结束后，有人问了线程迁移的相关工作，也有人问迁移的开销等等，而 Tong Li 的回答也都非常到位。总体上，他们的工作和报告都很出色。

第三篇文章是来自 Princeton 的 Ruby Lee 教授组里的工作，是关于安全方面的。他们把思想实现到了 OpenSparc 中，并在 FPGA 上运行，工作非常扎实。但很遗憾的是他们的报告很平淡甚至有点枯燥，听完后不知道他们到底在做什么，当然也很可能是我对这个领域的背景知识了解的太少了。

第四篇文章来头不小，是 CMU 的 Onur Mutlu 小组的工作，关于多个内存控制器调度问题。近几年 Onur Mutlu 在内存控制器调度方面发表了很多文章，尽管有一些学术界和工业界的研究人员对他们的工作不以为然，但不可否认的是，他已经成为体系结构领域最耀眼的新星了。其实，我们小组的研究方向和他的方向也有一些相关，所以我也比较关注他们的工作。他们工作中涉及到 Pareto 分布，提到已经有很多研究表明计算系统中许多事件都服从 pareto 分布，比如进程生命周期、Web 中传输的数据大小、文件大小等。他们也是首先假设访存请求服从 Pareto 分布，然后再去验证。然后根据 Pareto 分布特点选择了 Least-Attained-Service (LAS)调动策略，进而改进为 ATLAS（Adative per-Thread LAS）算法。总的来说，这篇文章从结构、实验到数据分析，写得很不错，这里就不多作介绍了，感兴趣的去看文章吧。报告是由一个韩国人作的，讲得也很清楚。但是下面观众抛给他的问题却是相当的尖锐，例如，他们开始提到新算法具有更好的扩展性，但是有一个图却显示随着内存控制器数目越大，他们的算法比其他方法性能提高的幅度越小。总的来说，我觉得他有个别问题回答的不是很好。

每个报告都有 6~8 个问题，进行了大约半个小时。这个 Best Paper Session 有两位中国人，而且不管是论文还是报告，都很精彩，这的确很鼓舞人心。所以，我也理所当然地投了他们的票。

五、其他报告

我在出发前询问过实验室其他人有没有特别想关注的文章。吕慧伟师弟做模拟器的，他让留意 MIT 的 Graphite 并行模拟器；二林是数学出身，想要关注 Onur Mutlu 他们那篇文章中访存请求服从 Pareto 分布的推导（前面我已经介绍）。

近几年，随着多核甚至众核盛行，并行模拟器也成为研究热点。比如 SC'09 上有文章介绍 P-Mambo，目标是在"Cluster-on-Cluster"；这届 HPCA 上 MIT 的 Graphite 目标是在"Multicore-on-Multicore"，要模拟一个 1000 核系统。并行模拟器的核心思想都是把目标机器分解为多个模块，这些模块可以分布到不同的处理器上同时运行，而模块之间通过消息传递数据。思想虽然很简单，但是作为一个系统，实现出来需要很大的工作量，而且每个系统都应该有一些自己的特色。Graphite 比较巧妙的地方是利用 Pin 来截获一些需要模拟的部分，比如访存、系统调用、网络消息等，而其他部分采用了 Direct Exection 模式。ASG 小组(http://asg.ict.ac.cn/en/)也开发了一个并行模拟器框架 SimK，已用于曙光中大规模网络模拟，现正在利用 SimK 框架将 Godson-T 众核处理器模拟器并行化。我一直觉得如果可以把 M5 用 SimK 并行化，也许会更有影响力。

紧跟着 Graphite 的报告是来自比利时 Davy Genbrugge 等的 Interval Simulation 工作。一开始我并没有注意到这篇文章，直到听了几页 ppt 才忽然意识到 Interval Simuation 我原来已经看过了，其核心思想就是 James Smith 他们发表在 TOCS 上的 OoO 处理器的新 Performance Model。他们认为 Interval simulation 的最大贡献将 Analysical Modeling 与 Cycle-accurate simulation 相结合，其实原理很容易理解，有点类似于 PDES 中的时钟同步思想，即在确定一个事件 Stall Cycles 的时候就直接大幅度推进模拟时间。Interval simulation 方法的效果非常好，把 M5 改造后与原来精确时钟模拟相比误差在 10%以内，但模拟时间可以降低 10 倍以上。我问了一个问题，Interval Simulation 对于固定 Stall Cycles 可以比较好的模拟 效果，但是对于一些 Stall Cycles 不确定的事件怎么处理，比如访问 DRAM 的 cycle 就会有很大的变化。Genbrugge 回答说他们以后会考虑这类更详细的模拟。后来我自己思考后觉得，他回答得很对，只要再把 DRAM 模型细化为更细粒度的事件，比如，ACTIVE、PRECHARGE、READ 等，那么就确定可以把这些事件的 Stall cycles 了。

Memory 功耗也是我所关心。现在很多数据都认为内存功耗占整个系统的 1/4 到 1/3，而降低功耗有很多种方法，比如降低电压，而学术界主要研究的方法有：1.关闭不访问的内存；2.优化调度算法。然后研究关闭哪些内存、什么时候关闭等。今年 HPCA 上 Mingsong Bi（大学本科同学）他们提出在内存被大量用于 buffer cache 时，可以利用系统调用信息提前预测打开和关闭哪些内存。再者，Micro'09 也有一篇利用 malloc 信息来有选择的 refresh 内存以降低内存功耗。

我还听了好几个中国人的报告，不过因为方向问题，有些工作了解地并不深入，这里也就不再做介绍了。

六、我们的报告

我们的报告在 12 日下午，是 Session 4B 的最后一个。Session 4B 是 On-chip Network & I/O，一共四个报告，有三个是中国人，分别是 Pissbutr 的 Yi Xu，Intel 的 Yaozhu Dong 和我（其中 Yi Xu 和我是本科系友；更神奇的是这次 HPCA 上有两篇一作文章的 Mingsong Bi 跟我本科同届，那时我们的学号分别是 3 号和 4 号，经常在一起做实验）。

12 日上午我又把 ppt 整理了一篇，中午吃完饭后早早赶到了会场。1 点刚过，会场里面的人陆续增多。一会儿 Session Chair PSU 的 Vijay Narayanan 也进来了，在安排我们测试投影

的同时，他也让现在所有的人都自我介绍一下，还风趣地说待会提问时要拷问报告人是否记得观众名字。现场原来肃静的气氛一下子活跃起来了，一个挨一个地介绍起来。我也感觉特别地轻松，似乎就像是在所里作个主题沙龙报告一样，只是少了一些水果。

前面的三个报告都顺利完成了，终于轮到我出场了。这个报告的中文版我在所里讲过两次，不过英文版还从没有讲过。一路讲下去，自己感觉到不算流畅，但观众应该能听懂。我在做报告 ppt 时主要侧重两点：一是通过访存 Trace 分析 I/O 数据与 CPU 数据访问特征不同，二是利用 Cache Global State 方法证明将 I/O 数据与 CPU 数据分开后引起的异构 Cache Coherence Protocol 无冲突。同时，我觉得还有些问题值得进一步研究，也在 ppt 中列了出来。比如，异构 Cache coherence Protocol 设计问题，如果 GPU 和 CPU 融入到一个地址空间，那么一致性协议该如何设计是一个必须面对的问题；考虑 I/O 因素的存储层次设计，特别是当 PCIe、SSD、GPU 等一些高速 I/O 总线和设备越来越流行，技术不断提高，传统的存储层次中 I/O 数据所占的比例在不断地增加。

提问阶段遇到了一个尴尬的场景。一个印度人问了一个问题，但是我真的没有听明白他到底问了什么。而 Vijay 试图帮我复述问题时，他突然也卡壳了，问那个印度人“what’s your question?”后来有一个后排的人似乎理解了这个问题，重新陈述了一遍，而 Intel 的 Yaozhu Dong 和 Jianhui Li 也帮我解释（真的很感谢这些朋友）。

我终于知道了他们的问题，主要是问真实应用的性能提高。这正是我们实验的一个不足之处，因为我们的实验评估指标是内存总线时钟数，而不是最终的应用性能。我想首先解释实验和评估方法，然后想再说明我们采用的 Trace-Driven 方法暂时还不能直接评估最终性能。实际上，如何利用 Trace-Driven 来评估存在多个层次事件 Overlap 情况下的最终性能，这是非常有意义而也有难度的问题。其实核心问题就是如何分析出不同层次事件 Overlap 的比例。ISCA2004 年有一篇关于 Memory Level Parallesim 的文章，曾经提出过一种方法计算 Compute 与 Memory Access 的 Overlap 比例与 MLP 值，但这只是一种通过统计加公式推导的方法得到的，还不适合细粒度的 Trace-Driven 场景。

七、Best Paper Award

13 日中午 12 点，本届 HPCA 的压轴戏就要上台了——公布 Best Paper Award。Program Chair Pradip Bose 详细介绍了选取 Best Paper 的流程，除了根据 Review 分数选出 4 篇候选文章供大家投票，还有一个 5 人的 Best Paper Committee 来为这 4 篇文章排名，最后比较大众投票结果和 committee 排名结果。

Pradip Bose 最后宣布：大众投票得出的第一名和 Committee 评出的第一名是同一个！那就是 Tong Li 等的“Operating System Support for Overlapping-ISA Heterogeneous Multicore Architectures”。很有意思，虽然 Keynoter 和 Panelist 们没人谈及 OS，但观众们显然还是对多核 OS 设计颇感兴趣的。看着 Tong Li 走上台去领奖，真替他高兴。而且 Tong Li 也是 PPoPP’09 的最佳论文获得者，真是佩服。再加上 PPoPP’10 的 Best Paper 由 Xipeng Shen 获得，今年中国人包揽了 HPCA 和 PPoPP 的最佳论文！

我们很荣幸地邀请的 Tong Li 博士在 2 月到北京来访问计算所和清华，他会介绍他们这个 OS 方面的工作，欢迎到时大家参加。（本来定在 2 月 3 日，所里的宣传海报都已经制作完成。但因为 Tong Li 临时有急事，所以要推迟几天。）

八、交流

8.1 室友——清华的翟季冬

这次很幸运和翟季冬结伴而行。他是清华陈文光老师的得意弟子，09 年发表了一篇 SC，10 年又发表了一篇 PPoPP，在高性能计算领域的研究工作开展得非常出色。一路上我们海阔天空地聊天，从各自的研究工作到各自实验室的情况，从计算所的状况到清华的研究，从自己的体会到一些八卦，从曙光到龙芯，从中国到印度……和他的交流感觉有很多共同语言，让我感到非常轻松而又收获颇多。我想，产生这种感觉可能的原因是，一方面我们的经历非常的相似，都一直在国内而没有海外求学的经历，对大陆的科研环境也都有着相同的体会；另一方面，我们来自不同的单位，对彼此单位都有些好奇，希望能了解更多。虽然以前我们见过面，相互认识，但在国内却从没有机会能深入交流，反而是这次 HPCA 与 PPoPP 合办创造了这样的机会。也许，国内的机构和企业可以考虑多创造些机会让国内的年轻人们聚聚。

翟季冬的工作是跟高性能计算相关，基于小规模节点预测大规模 MPI 程序的性能。我也邀请他回国后到计算所访问作报告。

8.2 Best Paper Award——Intel 的 Tong Li

11 号晚上的晚宴上结识了 Tong Li。从他的谈话中，能明显地感觉到他对研究有自己的理解，分析问题时思路很清晰。

我们谈到 MIT 的 Corey 以及 ASG 小组袁清波开发的 GenerOS，两者出发点有些类似，为了减少多核系统中锁和 Context Switch 的开销，把 OS 划分为多个模块绑定到专用核上，实验表明对 Oracle、Web 等应用有较好的性能提高。不过 Tong Li 质疑两点：

1）是锁的开销真的有那么大吗？他不认为锁的开销有那么严重，并且认为锁的开销可能只是一种表象。他们曾经在实验中也发现内核中有个锁开销很大，但后来发现是因为有个参数配置的问题，修改后就不存在那个锁开销就几乎没有了。对此，我的观点是，锁的开销取决于请求锁的线程数以及 Critical Section 的大小。可以考察两个极端，一个是如果能将所有 Critical Section 分解到很细的粒度，那么这种细粒度锁的开销肯定会显著降低，可以获得很好的并发性能，但是这样实现复杂度就会大大增加。另一个是像早期 Linux 使用 Big Kernel Lock，实现非常简单，但 Lock 的 Critical Section 非常大，导致性能非常差。因此，关于 Lock 的设计就是一个性能与实现的 Tradeoff，如果能消除锁并且达到较好的并发性能、同时还可以降低实现复杂度，那也是有意义的。

2）对于个别应用的提高真的就能是新结构比现在的 OS 结构好吗？不仅仅针对所谈的 OS 结构，他认为体系结构研究提出的新技术也一样，如果要面向市场，那么仅仅通过个别应用的实验数据是没有说服力的，需要有大量的应用做测试。在 Intel 内部，每推出一个新产品，都需要运行上千个应用进行测试。有的 Intel 部门向 Tong Li 所在部门推销新技术，号称有 10%的性能提高，但用更详细的模拟器后就只有 3%的性能提高了，接着再用上千个应用去测试时可能只有 30%的应用才有提高，而且还有一些会出现性能下降。对于这一点，我很认同。

我们聊得很投入，晚宴结束还觉得不过瘾，有一种相见恨晚的感觉。正好我们住在同一个酒店，于是 Tong Li 建议到我们房间继续聊。于是，Tong Li、翟季冬、Xin Sui（我们隔壁房间的，来自 UT Austin）和我齐聚我们房间，畅聊至凌晨 1 点半。

此时，我们聊的话题更广了，包括国内外科研现状、国内发展情况、国外的研究生活等。我们谈到了写文章的技巧。Tong Li 提到一篇文章如何写非常关键，国内很多工作其实做的非

常出色，但是就是不会包装。他举了去年那篇 PPoPP'09 最佳论文的例子，其实那篇文章之前投了 SOSP 被拒了，但最终一共有 8 个 Review，属于比较有争议的情况。Tong Li 于是重新调整了文章的结构，把 Review 认为比较晦涩难懂的理论分析部分挪到了文章后半段，投了 PPoPP，便获得了非常高的分。我提到，我们这篇 HPCA 文章在结构上做过 3 次重大调整，也是在不断地摸索中才慢慢领悟其中要领，所以对于写文章的重要性也有深刻体会。

Tong Li 在美国已经呆了十多年，已有很多年没有回国，所以非常渴望了解国内的情况。用他的话说，“在国外呆的时间越长，爱国心越强，就会很想回国来做点事；但对国内的现状又不是很了解，也很担心回来后不能如愿”。这的确是实话，应该代表了众多在海外闯荡的中国人的心声。国内科研的经费比以前充足，但是一些制度却并没有发生变化，因此屡有海龟水土不服的新闻报道。前段时间还看到 Yuanyuan Zhou 在 twitter 上表达对浙大海龟跳楼事件感到非常悲伤，这些事件必然会使想那些回国发展的海外学者们蒙上心理阴影。

我们也就是在这次聊天中邀请了 Tong Li 到北京访问。

8.3 Wen-Mei Hwu 和他的弟子 I-Jui Sung

之前提到 10 日清晨等车时遇见了 Wen-Mei Hwu 教授，在大巴上我们坐一起；后来在 Workshop 会场，他也坐在我旁边，所以有很多机会和他聊天。我问了 Wen-Mei Hwu 教授对编译开发自动并行的看法。他认为，仅仅靠编译来实现自动并行显然不够了，需要加入一些制导，这就好比造房子需要脚手架一样。Wen-Mei Hwu 还提到，自动并行比较难处理的是指针问题，还有就是数据结构变换问题。对于指针问题，后来和翟季冬交流时，他也对此有同样的见解，指针的 point-to 问题是编译中的经典问题，却也是老大难问题，全世界研究了很多年也没有什么进展；而这个问题又是很多编译优化的基础，包括他们现在的工作也因为这个指针 point-to 问题，而暂时对 Fortran 处理的比较好，而对 C 语言处理还有不足。

Wen-Mei Hwu 还介绍他的学生 I-Jui Sung（宋懿瑞），让我们彼此认识。I-Jui 是一个很随和的人，12 日晚上坐大巴去 Bangalore Palace 参加晚宴的，我和他坐在一起，交流了一些大陆和台湾的 IT 状况。从他口中得知，台湾学生一般更喜欢读 EE 而不是 CS，因为 EE 更容易就业；甚至在学校里，EE 的地位都比 CS 高，比如他在交大 CS 是第一名，奖学金只有 3000 新台币，但是 EE 的奖学金却有 30000 新台币；最近几年联发科的崛起稍稍改变了这种局面，因为联发科因为不重视软件而栽过跟头；他发现到了美国后，情况急转掉头，CS 比 EE 红多了，美国学校 CS 的经费要比 EE 多很多，而工作也更好找。

我们也聊起了 Wen-Mei Hwu 如何培养学生，I-Jui 说，

1）Wen-Mei 并不会具体指导他们做什么，他一般会给学生几个应用，让学生熟悉如何优化，包括手工优化，然后再深入研究编译优化；Wen-Mei 的关系网太强大了，似乎什么人都认识，几乎每周都会邀请其他大学或者工业界的人来作报告；

2）此外还会安排小组成员与受邀学者面谈，每个人介绍自己的工作，如果有该学者现在做的相关，那么可以进一步考虑合作，如果是企业嘉宾，那么就会派学生去实习。

8.4 CMU 的 Onur Mutlu

在 11 日我就先和 Onur 聊过一次，介绍了我们也在做 Memory System 相关方面的研究，并介绍 HMTT Trace，如果有机会可以一起合作。他说非常有趣，要听我们的报告，问报告的具体时间。12 日下午，果然在我们报告开始之前，我看到他进入会场。报告结束后，他跟我说，我们提出的分离 I/O 与 CPU 数据概念非常有趣，我们可以谈谈合作的事情。于是，我们约好了在 Bangalore Palace 的晚宴上聊聊。

晚宴上，我们一起合了个影，希望这是一个好的合作起点。我们首先聊了一下 HMTT

的基本原理、访存 Trace 所提供的信息和 Benchmark。我带过去两张刻有 8GB Memory Trace 片段（四个应用 netperf、tpc-c、specjbb 和 specweb）的 DVD 也给了他。接着我跟他谈了我们现在的一些想法，他觉得很好，非常乐意一起研究。于是我们开始商量合作方法，我本来想直接派一个学生到他组里访问三个月，但和组长陈老师商量综合考虑后，第一阶段还是决定先以邮件电话交流为主，开展一些头脑风暴。

我们也侃了一些其他有趣的事，比如职称体系。我提到，虽然我也是 Assistant Professor，却并不能招收学生，而要带博士生更是许多年以后的事情。对此，他也是觉得比较惊讶，问我小组的学生是哪来的？"The Big Boss!……"大家都笑了。

图 7. 与 Onur 的合影

8.5 其他交流

开会这几天和很多人交流过，很多都是蜻蜓点水般聊上几句而已，不过也有一些也让我很有收获。

- **Pen-chung Yew 教授**：Yew 教授和国内很大研究机构和大学都有合作。在我报告结束后，他就和我聊起了 HMTT 系统，和我一起分析该如何回答那个印度人问的关于实验评估问题。他说印度人的口音是比较重，教我以后遇到这种情况，可以说“let's discuss offline”。在 Bangalore Palace 的晚宴上，由于演出的音响非常震耳，Yew 教授一个人坐到了最后。国防科大的 Libo Huang 和我也坐到他旁边，大家聊了一些美国、大陆和台湾的计算机领域的研究模式。其实在 Yew 教授所在的 Minnesota 大学也是以申请国家或者军方的项目来驱动研究，不过他们的项目都比较有挑战性，因此也比较容易出成果。
- **Yuan Xie 教授与他的学生 Guangyu Sun**：Xie 教授是 HPCA 的 Student Affairs Chair，负责向申请 Travel Grant 的学生提供经费。他说，一般为一作提供 1200 美元经费，二作是 800 美元，而没有文章的学生如果申请的话也能获得一些经费。一个有趣的小插曲，翟季冬说他忘了申请 Travel Grant 了，Xie 教授说没事，你现在给我发一封邮件就行了。季冬非常高兴，说为陈老师省钱了，忽然想起一个问题，便问道：“我是参加 PPoPP 也可以申请 HPCA 的 Travel Grant 吗？”“啊！这个不行，我只能管 HPCA 的学生。”我们都笑了，感到一丝可惜，但心里很感谢 Xie 教授那份好意。大家谈到考核问题，Xie 教授说，在他们学校里面，评 Tenure 主要看三点：1）研究，包括影响力、文章、项目等；2）教学，主要是由学生打分；3）学术活动的服务，比如他这次作的 Student Affairs Chair。

  Guangyu是Xie教授的学生，他有一篇用PCM改善SSD性能与功耗的文章。我们一起探讨了一些新颖的内存技术，前段时间正好看到了一篇评论文章，分析了五种可能有重大革新的内存新技术优缺点（分别是MRAM、FeRAM、PCRAM、RRAM和Racetrack Memory），我们讨论地饶有兴趣。（感兴趣的朋友可以看这篇文章 “Five ways to revolutionise computer memory”， http://www.newscientist.com/article/mg20427370.600）

- **Arvind 教授**：在 Bangalore Palace 的晚宴上，我和 Arvind 聊了一小会。我问，在他的 Keynote 中提到每个 Core 都负责一种专用功能，那么如果多个 Core 之间要共享数据，会不会对 Cache Coherence 设计带来挑战。他认为，肯定会对 Cache Coherence 设计有一些新的挑战；但如果是在 Manycore 系统中，特别是千核规模，还是需要软件层次降低对 Cache Coherence 的依赖，否则，Coherence 的开销会非常大。我们也提到 2008 年的 ISCA，他认为办得非常棒，又问我是否认识 Xiaowei Shen。
- **Xipeng Shen 教授**：13 日上午是翟季冬的报告，我去 PPoPP 会场旁听，并帮他录像。我看 Xipeng Shen 旁边有空位，于是坐了下来。Xipeng Shen 他们的文章是今年的 PPoPP 最佳论文，我没有听到他们的报告，于是问他要了 slides。

上次他到计算所介绍了“Input-Centric Program Behavior Analysis”，我觉得很有趣，于是继续向他请教。我觉得，程序的 Input 是一个非常大的空间，而程序第一步都会根据输入的参数在这个大的 Input 空间定位到一个小的子空间，然后再做处理。那么有没有技术可以通过分析程序代码就可以把一个完整的 Input 大空间划分为多个子空间呢？比如，一个程序判断输入 n<100 时做一种处理，而 n>=100 时又做另外一种处理，那么就可以通过分析程序把 input 空间划分为两个子空间[0 100)和[100, inf)。我觉得这个工作应该很有意义，比如 Benchmak 设计时选择什么作为输入集合呢？Xipeng 也比较认同这个观点，他觉得根据源码分析应该是可以能做到这一点的，但难度也不小。

- **MIT 的 Jason Miller 博士后**：Jason 就是做 Graphite 并行模拟器的。我跟他聊了并行模拟器设计的一些问题，比如吕慧伟师弟问的如何做到在时钟不精确的模拟器上去研究 Coherence 问题。他认为这要看 Coherence 的层次，如果是应用级别的，那是没有问题的；但如果是硬件层次的 Cache Coherence，那么就需要时钟精确级别的模拟器了。

他们还有一个 PPoPP Poster，题目很吸引人“Application Heartbeats for Software Performance and Health”，通过 Binary Instrument 技术对应用程序的一些关键步骤插入产生 Hearbeat 的代码，这样就可以实时监控程序的运行状态了。虽然感觉思想和实现都不复杂，但我觉得这个工作做好了很有意义，很多时候我们把程序启动后，运行了几个小时没有动静，也不知道它运行的怎么样了？有没有陷入死循环？。

图 8. 与 MIT 的 Jason 合影

九、中国人的聚会

大概有 20 多个中国人参加了这届 HPCA/PPoPP，人数应该仅次于印度和美国。

能在中美之外印度聚起一群中国人，一起逛街、购物、吃饭，真的非常难得。在会议结束后的那个下午（13 日），大家自由活动。一部分 Intel 的朋友去拜访 Bangalore 的 Intel 分部，国防科大的黄立波和老师他们另有安排，一些大教授们也是自行活动。剩下还有我们 9 个中国人，加上几个欧洲人，在一个印度学生的带领下，浩浩荡荡前往 Bangalore 最繁华的商业

街购物，甚是壮观。

图 9. 一部分中国人合影

在商业街购物，砍价自然必不可少。NCSU 的 Fang Liu 看中了一个挂袋，老板开价600卢比（可能更高），她想砍到300；老板不卖，于是 University of Pittsburgh 的 Yi Xu 凑上去也要买，2个600；老板还是不卖，我也凑上去，3个900；依然不卖，UIUC 的 Lei Chen 也凑了上来，4个1200。很遗憾，老板依然不卖，但这种仗着人多势众讨价还价非常有趣。这次购物很有收获，我和 Yan Pan 两个买了印度地图，我还买了印度丝巾，Xipeng Shen 对印度木雕情有独钟，翟季冬和 Xin Sui 都买了1200卢比一个的木雕象……

晚上我们九个人来到 Xipeng Shen、Fang Liu、Yi Xu 和 Yi Guo 他们住的酒店一起吃自助，恰好看到 Intel 的 Yaozhu Dong 和 Jianhui Li 也在餐厅。有段时间，整个餐厅就我们11个人在就餐。自助的味道一般，但是一群人七嘴八舌聊得还是很愉快。

Intel 的 Jianhui Li 在回国后还专门建了一个所有参会中国人的邮件列表，很是用心。这次聚会给我印象很深刻。

十、Bangalore 印象及自助一日游

在飞机上，翟季冬跟我提到他同学去新德里的糟糕经历，酒店条件非常差，所以提醒一定要自备饮水和食物。于是，我的行李中便多了 3 升水和 5 袋方便面。当 9 日晚上 10 点多和翟季冬一起到达 Bangalore 机场，感觉机场大厅很现代化、也很整洁。我们抵达酒店，发现条件相当的好。Bangalore 给我的第一印象还不错。

10 日下午，我和翟季冬从会场走回酒店，大约 4 公里路，花了 1 个多小时，领略了 Bangalore 的街道。Bangalore 被誉为“印度硅谷”，是印度最现代的城市。但 Bangalore 市区的基础设施估计与国内 90 年代初的东部沿海县城相当，还是比较破旧。大街上主要交通工具是摩托车，摩的则是他们最流行的出租车，公共汽车基本都没有门，经常看到乘客半个身体掉在车外，也许这样比较凉快。这里的路标比较混乱，有的路口都没有标识，所以也导致我们根据 google 查好的路线却因为找不到路口标识而走了很多冤枉路。回到酒店，我们上网查了两个地点，市政大楼 Vidhana Soudha 和红房子高等法院 Attara Kacheri，再次步行前往。再加上 Bangalore Palace，这几个建设都是比较有印度特色的，也都有历史（具体大家去查阅 wiki 吧）。

图10. Lal Bagh 公园中佛塔在积聚祥云

图11. 马路上的印度神牛

图12. Bangalore Palace

图13. Bangalore 街头

14日，大部队都去400公里外的 Belur 和 Halebid 参观寺庙了。我因为要赶晚上的飞机，就决定在市里自助一日游。我叫了一辆摩的，事前谈好价钱，一天下来也转了好几个地方：中国餐馆、Lal Bagh、Bull Template 和 Science & Technology Museum。

Bangalore 当地人还是很友善，估计很少见到中国人，所以在 Lal Bagh 公园的时候，有好几个印度人都邀请我和他们合影。我特地去了一趟 Science & Technology Museum，相当不错。价格只要20卢比（约人民币3元），博物馆不大，主要是侧重机械设备，比如1：1仿制了莱特兄弟的第一架飞机、飞机发动机、火箭整流罩等。所有设施都是开放的，游客可以自己去体验。我看到有很多的印度学生都在摆弄，看起来很是好学。

图14. 印度游客邀请和他合影

图15.印度学生在摆弄设备

不过，还是有几个地方需要注意：

1. 事前要把去的地方定好，否则那个司机最后要加钱。我本来和那个司机说好是 300 元，但后来加了两个地方，也多加了 150 元；

2. 那些司机会带我们到一些店去买东西，和这些司机聊天才知道这些店是他们出租车公司开的。如果司机带旅客过去，达到一定人数就可以得到免费汽油。可以不买。

3. 记得拿个地图，你自己需要先把地点与路线大概了解一下，否则这些司机总会说不知道在哪里。他们总是想尽快把旅客带到商店去。我幸好在酒店拿了一个地图，然后我把手机上的 GPS 和 Google 地图数据都更新了一下，并把路线也搜索出来了保存下来。所以司机说不认识的时候，我就拿手机给他看，结果他就不磨蹭了；

4. 当要去一个地方的时候，司机经常会说那个地方关门的。不要听他们的，他们是在唬人，坚持一定要去。

5. 在 Bull Template 时，有一个人跟我说这里不能乱走，要按照一套仪式走。于是他便带着我走了一遍，结果出去的时候却问我要 30 美元。我感觉就是上当了，就跟他理论，最后给了 6 美元。所以，遇到这种情况要千万小心。

总的来说，这些司机都很精明，需要自己先做点功课，多查一些资料。不过，Meru 出租车公司倒是比较正规的，我约了一辆车，司机非常好。

十二、小建议

1. 国内出去开会，背景了解会少一些，所以开会前最好要做一些准备工作。把可能出现的大牛资料都查阅一遍，比如照片、研究精华、哪些学生也参加会议，这样会提高搭讪的效率，增加印象。我觉得这次出门前准备工作做得不好，导致有些人聊过以后都没有印象了。

2. 国内出去开会能和其他单位的同行结伴出行是非常好的选择。这是小同行深入交流的好机会。同时，也希望国内能多创造一些机让一些博士生和刚毕业的博士交流的机会。

3. 带一个电力充足的相机或者 DV，这样可以把 Keynote 和 Panel 给拍摄下来。有很多内容现场听了以后消化不了，回来后可以再复习消化一下，更有收获。