# 个人陈述

## 1 申请动机

攻读硕士期间，我的主要研究领域是光纤通信系统。由于光通信系统容量日益增长，需要复杂的数字信号处理算法来弥补高速传输带来的物理损伤，这也使得 DSP 的功耗上升，成本增加。而当前的光网络节点数量日益增长，无法承受过高的单链路成本。在这个背景下，我的主要研究方向是光通信系统中的低复杂度数字信号处理算法。随着研究的深入，我逐渐认识到数字信号处理器（DSP）是支撑光通信系统发展的基石 [18]。与此同时，我也逐渐感受到自己研究工作的局限性：我所有的工作都是基于仿真或者实验获取接收机的数字信号，然后使用 Python 和 MATLAB 在实验室电脑上进行离线处理，但我从没验证过算法在 DSP 中的实际效果，甚至完全不清楚 DSP 硬件的架构，更不知道其如何执行所研究的算法。我开始怀疑自己使用通用计算机的 CPU 跑 MATLAB 程序得到的结论是否与 DSP 中的实际情况存在出入。这样的疑惑让我对不同处理器的架构，以及一般意义上的计算机系统与体系结构（下称 Sys-Arch）产生了兴趣，而我的两段研究经历让这种兴趣变得更加强烈。

（1）在研究基于 Stokes 空间解偏振 [23] 的均衡算法时，我了解到这种算法相比于传统的 CMA 算法 [24] 具有更好的鲁棒性和更小的训练开销。这让我不禁好奇：为何工业界采用的是 CMA 算法？与通用计算机不同，DSP 是一种处理流式数据的实时系统，而 CMA 算法每次迭代只依赖当前时间窗口内的信号，恰好与 DSP 系统匹配。尽管可能需要上万个信号才能收敛，但内存中只需要保存一个时间窗口内的信号（最多几十个）。相比之下，基于 Stokes 空间解偏振的算法虽然只需要上千个信号，但受到算法原理限制，需要依赖所有的信号才能计算出抽头系数，因此这些信号可能都需要加载到内存中。此外，DSP 还针对 CMA 这种迭代式的计算有专门的优化（software pipeline 等），因此我猜测 CMA 算法在实际硬件中，可能内存开销和延迟都更加优秀。站在通信的角度，Stokes 算法需要更少的训练信号从而 “冗余” 更少，但仅仅从这个角度就说其比 CMA 算法有更小的 “开销” 却是片面的。

（2）我曾经调研过一段时间所谓的 “光计算”—— 随着深度学习的爆发，针对深度学习训练和推理的专用硬件的研究也延伸到了光学领域。光集成器件具有天然的并行处理能力，并且有极低的功耗。基于此，学术界 demo 了很多光子神经网络 [20, 15, 27]。目前光学领域，大家主要在针对光子神经网络推理精度和如何利用光器件实现非线性运算进行研究，我也相信随着光子技术的发展，这些问题可能得到有效的解决。但是，我也时常在思考另一个问题：这些光集成电路如何才能变成一个真正可用的计算机系统？尽管在神经网络推理的场景下光计算有望超越电子计算机，但是在可见的未来光计算无法实现绝大多数通用计算机的功能。因此，用于加速神经网络推理的 “光处理器”（OPU）必然需要作为现代计算机的协处理器存在——类似于 GPU 和 CPU。要实现这样的系统，就需要将 CPU 内存中的数据 “喂” 给 OPU。对于一个 CPU-GPU 的系统，通过 PCIe/NVLink 可以将数据传输到 GPU 的显存中，但是目前为止光集成电路并不能实现 “内存” 的功能，因此数据只能实时传输给 OPU 进行运算，然后再实时传输回 CPU。此外，由于 OPU 只能处理模拟光信号，这个 copy-in 和 copy-back 的通路上还需要高速的模数转换和光电转换。如此一来，整个系统可能存在较大的数据传输开销，且需要承受更高的硬件成本，OPU 的加速性能是否还能体现呢？

两段研究经历带给了我许多启发。对于 (1) 中的信号处理问题，实际上是因为 CMA 算法是一种通用的信号处理算法，因此 DSP 硬件针对这种通用的软件特性做了优化。换言之，DSP 可以看作是一种 domain-specific 的 accelerator。**在这个计算需求复杂且多样的时代，针对不同的领域进行软硬件的协同设计价值重大**。(2) 中的光计算问题，本质上是一个异构计算的问题。随着物理器件的发展，新兴的计算范式例如光计算、量子计算等得以实现。而要真正体现这些计算能力的优势，则需要实现一个自底向上的异构计算平台。**Sys-Arch 领域的研究，正是连接器件与上层应用的桥梁**。以上经历和思考便是我申请博士的动力来源。

## 2 学科背景

目前我在 “信号与信息处理”、“计算机应用技术” 两个领域积累的知识、技能、研究经验如下。

**信号与信息处理：**在物理层通信技术里，我了解常见的编码、调制格式以及在发射机和接收机中常用的信号处理算法。我主要研究了基于 FIR 的数字均衡算法，了解基于 CMA，MMA，LMS 的抽头系数寻优方法。另外我还研究了

光纤通信系统中特有的解偏振算法，了解基于 FIR 蝶形滤波器和 CMA/MMA 的解偏振算法，基于 Stokes 空间的解偏振算法，以及基于 ICA 的解偏振算法。我在北京邮电大学的徐坤老师和于振明老师的指导下进行光通信系统中的信号处理算法研究并发表了两篇期刊论文。此外我也有一些一般意义上的信息处理技术的应用和研究经验，我了解 GBDT、LR、HMM 等传统机器学习算法，也了解 MLP、CNN、RNN、Transformer 等基于神经网络的深度学习算法。在百度实习和工作期间，我有一些语音识别、文本分类等方面的模型应用经验。

**计算机应用技术：**我没有系统的计算机科学基础的课程学习经历，但是我通过自学一些经典教材掌握了部分基础知识。我主要参考《算法导论》[7] 学习了数据结构和算法；主要参考《操作系统概念》[10] 学习了操作系统原理；主要参考《计算机体系结构：一种量化方法》[11] 学习了计算机体系结构。此外，我在匹兹堡大学的Xulong Tang助理教授的指导下学习了计算机体系结构研究生课程，并获得了 A 的成绩。我在 Xulong Tang，Jun Yang和Youtao Zhang几位老师的指导下参与了“量子计算模拟器的 GPU 加速”这个研究项目并形成论文，目前被 HPCA-2022 接收[1]。在研究中，我积累了一定的 GPU 编程能力，能使用 vtune/nvprof/nsight 等工具进行程序性能分析，熟悉了 OpenMP/thrust/CUDA 等框架。我在百度工作期间主要负责搜索引擎相关的策略/算法研发，期间我学习了搜索引擎的基本原理，了解了 Bigtable，MapReduce 等大数据处理框架的原理。

一个大的领域里面，许多技术的思想都是相通的。在信息处理领域，对于 FIR 滤波器，我们利用梯度下降去寻找最优的抽头系数，在神经网络中同样利用梯度下降去寻找最优的权重。Volterra 均衡器之于线性均衡器，很像 FM 之于 LR，本质上都是通过特征组合来拟合非线性关系。对于计算机系统也是一样，操作系统对于进程/线程的管理调度和体系结构中的流水线指令执行也十分相似，本质上都是在提升 throughput。搜索引擎的架构设计和计算机系统中的内存层次结构也很相似——都是数据存储能力逐层递减，数据访问效率逐层增加。我的浅薄理解是，在泛信息处理领域，好的工作往往来源于数理统计上的直觉，而在泛计算机系统领域，更多的是一种设计和管理的艺术，在不同的应用驱动下寻找最优的 tradeoff。接触不同领域之后，我也发现自己的兴趣点更倾向于后者，这也是我申请的动力之一。通信、电子、计算机甚至物理学的一些基础知识我都有所掌握，但是缺乏全面、深入的学习和应用。我相信扎实的基础是做出 solid 的研究工作的必要条件，也会在今后的学习和研究工作中持续、深入地学习。

# 3 研究计划

**短期目标：**Sys-Arch 的设计根本上是要为上层的软件应用提供支撑。在如今这个计算需求复杂且多样化的时代，针对不同领域的应用特点进行 domain-specifc 的 acceleration 具有重要的意义和广阔的研究前景。近年来，针对不同应用场景做加速的顶尖研究层出不穷。例如神经网络训练/推理 [25, 2, 16]，生物信息学 [26, 12, 4]，计算机网络 [8]，量子计算 [14, 13]，或者是针对应用场景广泛的某一类算法进行优化 [5]。我近期参与的量子计算模拟器的 GPU 优化工作 [1]，同样也是针对 quantum simulator 这样一个特殊应用进行的加速。可以说，只要是有高性能计算需求的场景，就需要 Sys-Arch 针对性的优化。在攻读博士期间，我希望首先从 domain-specific acceleration 入手 Sys-Arch 的研究。首先明确一个应用场景，然后在这个场景下对 SOTA 的 Sys-Arch 的支持进行性能分析，发现存在的问题，明确优化方向。接着对场景下的应用分析其所用的数据结构和算法的特点，然后针对性能瓶颈进行优化设计。这一类研究，往往是通过已有的方法解决一个新的问题，也不失为不错的研究工作。最重要的是，我希望通过这段“初级”的研究训练深入的理解和运用 Sys-Arch 中经典的思想和方法，成为一个基础扎实，比较有经验的 Sys-Arch 领域的研究者。

**长期目标：**在高年级博士以及之后的研究中我希望涉足“基于新兴器件的异构计算机系统研究”。随着晶体管尺寸逐步逼近物理极限，长期来看，计算机性能的持续提升依赖于物理器件的突破。Sys-Arch 的创新一直都能对器件级进步带来的性能提升进行数量级的放大，我相信这一事实将延续下去。近年来，新兴的器件在一般性或者特定的场景下呈现出令人遐想的应用前景，基于此的 Sys-Arch 研究工作也层出不穷。例如基于非易失性存储器的内存内计算 [6]；基于忆阻器 [3, 19]、硅基光集成器件 [21, 22] 等的神经网络加速；基于超导 [9]、离子阱 [17] 等的量子计算机架构。我个人的理解是，要做好这样的研究，不仅需要对经典的 Sys-Arch 的理论十分熟悉，也需要一定的 physics insight，理解物理上的原理和限制。例如在 [9] 中，作者充分了解量子电路运行所需的复杂的模拟波形控制，才设计出了具有精确时间控制的超导量子计算机微架构原型。在我看来，这些基于新型器件的 Sys-Arch，很长时间内都只能作为计算机中的某

[1] 我原计划 2021 年秋季入学匹兹堡大学计算机系攻读 PhD，但是因为10043遭到拒签，遂改变计划申请国内博士。

个计算单元存在。当前的计算机中通常搭载了 CPU 和 GPU，未来可能还会有更多的执行特定任务的 *PU。也许 “异构” 将成为未来的计算机系统的标配特性，在这样的系统中，我想有几个问题值得长期探索：（1）如何设计经典/新兴计算单元之间的 Interface，使得不同计算单元之间的数据格式转换、数据传输带来的性能开销尽可能的小；（2）如何设计系统中的控制单元，对不同计算单元上执行的任务进行高效的管理；（3）如何构建完整的异构计算生态系统，从设计和验证新架构的模拟器到指令集，编译器，再到操作系统，每一个环节都有大量的空间去探索。如果说我的 “短期目标” 主要还是依赖于已有的方法和思想，“长期目标” 的实现则更依赖敏锐的洞察力和创新能力。未来希望自己能在研究中不断学习，逐步实现自己的研究计划。

# References


[1] Yilun Zhao et al. “Q-GPU: A Recipe of Optimizations for Quantum Circuit Simulation Using GPUs (Accepted)”. In: *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE. 2022.

[2] Manoj Alwani et al. “Fused-layer CNN accelerators”. In: *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE. 2016, pp. 1–12.

[3] Aayush Ankit et al. “PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference”. In: *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. 2019, pp. 715–731.

[4] Mau-Chung Frank Chang et al. “The smem seeding acceleration for dna sequence alignment”. In: *2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE. 2016, pp. 32–39.

[5] Xuhao Chen et al. “FlexMiner: A Pattern-Aware Accelerator for Graph Pattern Mining”. In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE. 2021, pp. 581–594.

[6] Ping Chi et al. “Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory”. In: *ACM SIGARCH Computer Architecture News* 44.3 (2016), pp. 27–39.

[7] Thomas H Cormen et al. *Introduction to algorithms*. MIT press, 2009.

[8] Salvatore Di Girolamo et al. “A RISC-V in-network accelerator for flexible high-performance low-power packet processing”. In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE. 2021, pp. 958–971.

[9] Xiang Fu et al. “An experimental microarchitecture for a superconducting quantum processor”. In: *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*. 2017, pp. 813–825.

[10] Peter B Galvin, Greg Gagne, Abraham Silberschatz, et al. *Operating system concepts*. Vol. 10. John Wiley & Sons, 2003.

[11] John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2011.

[12] Lei Jiang and Farzaneh Zokaee. “EXMA: A Genomics Accelerator for Exact-Matching”. In: *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE. 2021, pp. 399–411.

[13] Ang Li et al. “Density matrix quantum circuit simulation via the BSP machine on modern GPU clusters”. In: *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE. 2020, pp. 1–15.

[14] Gushu Li, Yunong Shi, and Ali Javadi-Abhari. “Software-Hardware Co-Optimization for Computational Chemistry on Superconducting Quantum Processors”. In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE. 2021, pp. 832–845.

[15] Xing Lin et al. "All-optical machine learning using diffractive deep neural networks". In: *Science* 361.6406 (2018), pp. 1004–1008.

[16] Shaoli Liu et al. "Cambricon: An instruction set architecture for neural networks". In: *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*. IEEE. 2016, pp. 393–405.

[17] Prakash Murali et al. "Architecting noisy intermediate-scale trapped ion quantum computers". In: *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. IEEE. 2020, pp. 529–542.

[18] Seb J Savory. "Digital coherent optical receivers: Algorithms and subsystems". In: *IEEE Journal of selected topics in quantum electronics* 16.5 (2010), pp. 1164–1179.

[19] Ali Shafiee et al. "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars". In: *ACM SIGARCH Computer Architecture News* 44.3 (2016), pp. 14–26.

[20] Yichen Shen et al. "Deep learning with coherent nanophotonic circuits". In: *Nature Photonics* 11.7 (2017), pp. 441–446.

[21] Kyle Shiflett et al. "Albireo: Energy-Efficient Acceleration of Convolutional Neural Networks via Silicon Photonics". In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE. 2021, pp. 860–873.

[22] Kyle Shiflett et al. "PIXEL: Photonic neural network accelerator". In: *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE. 2020, pp. 474–487.

[23] Bogdan Szafraniec, Bernd Nebendahl, and Todd Marshall. "Polarization demultiplexing in Stokes space". In: *Optics express* 18.17 (2010), pp. 17928–17939.

[24] A-J Van Der Veen and Arogyaswami Paulraj. "An analytical constant modulus algorithm". In: *IEEE Transactions on Signal Processing* 44.5 (1996), pp. 1136–1155.

[25] Swagath Venkataramani et al. "RaPiD: AI accelerator for ultra-low precision training and inference". In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE. 2021, pp. 153–166.

[26] Lingxi Wu et al. "Sieve: Scalable In-situ DRAM-based Accelerator Designs for Massively Parallel k-mer Matching". In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE. 2021, pp. 251–264.

[27] Xingyuan Xu et al. "11 TOPS photonic convolutional accelerator for optical neural networks". In: *Nature* 589.7840 (2021), pp. 44–51.