芯来科技 NUCLEI

AI专用RISC-V CPU内核构与自定义指令集扩展

芯来科技 马越

AI技术持续进步，应用需求端稳步增长，各行业已开始广泛采用AI技术

芯来科技 NUCLEI

AI技术进步

机器学习、自然语言处理和计算机视觉等技术的进步，造就了更复杂的AI应用。这扩大了AI的潜在使用范围，并增加了需求

与物联网的集成

AI与物联网（IoT）的集成变得更加普遍，推动了对能够处理和分析大量连接设备数据的AI解决方案的需求

云AI服务

云端AI服务的兴起使得AI对中小企业更加普及，增加了不同规模企业对AI的采用

大模型的崛起

大模型性能和效率的提升，使其在各个领域得到了广泛的应用；同时大模型的部署不再局限于云端，本地设备上的轻量化大模型可以更好的满足实时处理和低延迟应用

芯片半导体的发展对AI起到了关键性的作用

芯来科技
NUCLEI

为AI提供了强大的计算基础设施

推动了AI技术的进步和广泛应用

提升了AI系统的性能、效率和普及度

提升计算能力

先进的半导体技术，如更小的工艺节点（例如7nm、5nm和3nm），
使芯片具有更高的计算能力和效率，从而加速AI算法的训练和推理过程

专用AI芯片

这类芯片专门优化了机器学习和深度学习任务，显著提高了AI系统的性能和效率

降低能耗

新一代芯片在提高计算能力的同时，大大降低了训练和推理过程中的能耗，
有助于在资源受限的环境中实现AI应用，降低成本

边缘计算

高性能的AI计算能力可以嵌入到较小的设备中，如智能手机、物联网设备和汽车中，
推动了边缘计算的发展，使得AI可以在本地设备上进行实时处理，减少了对云计算资源的依赖，
提高了响应速度和数据隐私

创新加速

更强大的计算能力和更高效的硬件使得研究人员和开发者能够探索更复杂的模型和更广泛的应用场景，
从而推动AI技术的持续进步

CPU是AI芯片中必不可少的部分

芯来科技 NUCLEI

- **对于AI基础设施来说，CPU是必不可少的存在**
  - 高端AI服务器中每8个GPU需要至少搭载2个CPU
- **异构计算离不开CPU的加持**
  - CPU + GPU
  - CPU + FPGA
  - CPU + TPU
  - …
- **CPU在AI芯片中起到的作用**
  - 负责整个服务器的运算与控制，实现高效的资源分配和任务调度
  - 具有灵活性和通用性，可以完成GPU无法完成的数据预处理以及串行计算
  - 软件兼容性上拥有极大的优势
  - 启动维护、完全保护
  - 在许多细分场景下，用异构CPU架构做并行计算更高效、成本更低

X86

SSE: Streaming SIMD Extensions, 128b
AVX: Advanced Vector Extensions, 512b

arm

NEON: 64/128b
SVE+SME: Scalable Vector/Matrix Extension, 128-2048b

RISC-V顺应了行业大趋势——底层协议开放标准化

芯来科技 NUCLEI

维护 管理 发展

指令集(架构)

RISC-V是世界上第一个**高质量、开放标准的CPU指令集**

RISC-V适用于**所有计算系统：MCU-汽车电子-数据中心-超算**

RISC-V开放的架构，极大**加速了生态的发展**

开放标准的指令集 ≠ 免费的CPU Core IP

RISC-V基金会

RISC-V基金会是非营利性组织，落户**瑞士中立地区**

近**4000基金会成员**，来自大陆遍布欧美的70多个国家

RISC-V已经和ARM、X86呈三分天下之势，开放的CPU标准是未来的趋势

芯来科技 NUCLEI

三分天下之势

私有化
绝对封闭

X86

intel AMD

输出CPU IP授权
相对封闭

arm

公开化的国际标准

RISC-V → **顺势而生**

开放性 先进性
简洁性 模块化 扩展性

逐渐淡出主流

MIPS

SPARC

Sun

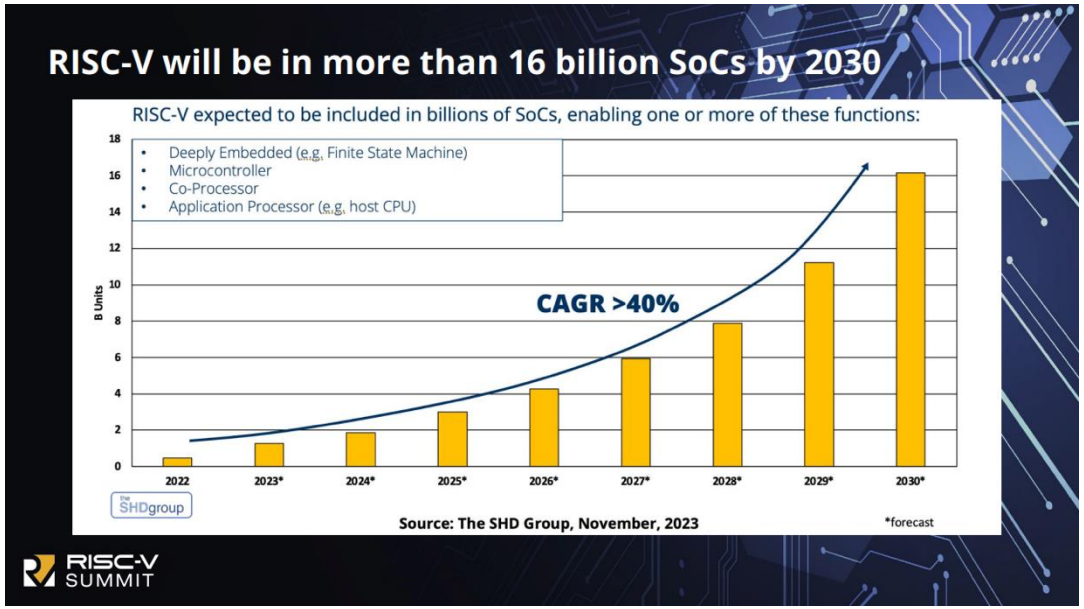
PowerPC

IBM

1978 1984 1986 2015

RISC-V未来十年增长迅猛，将覆盖各类应用场景

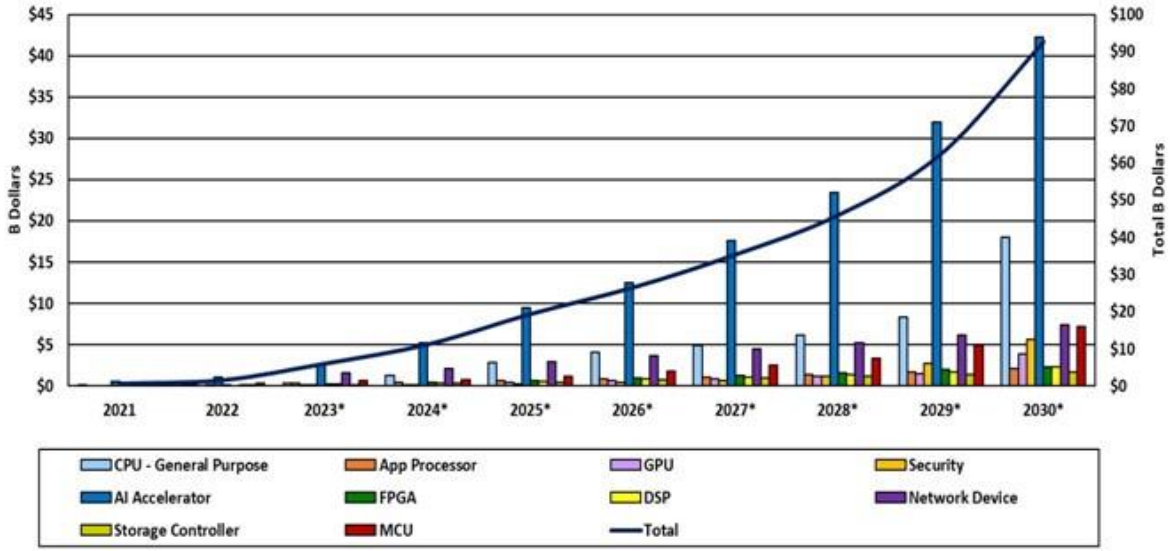
芯来科技 NUCLEI



RISC-V预计2030年出货量达到**160亿颗**
CAGR将超过**40%**

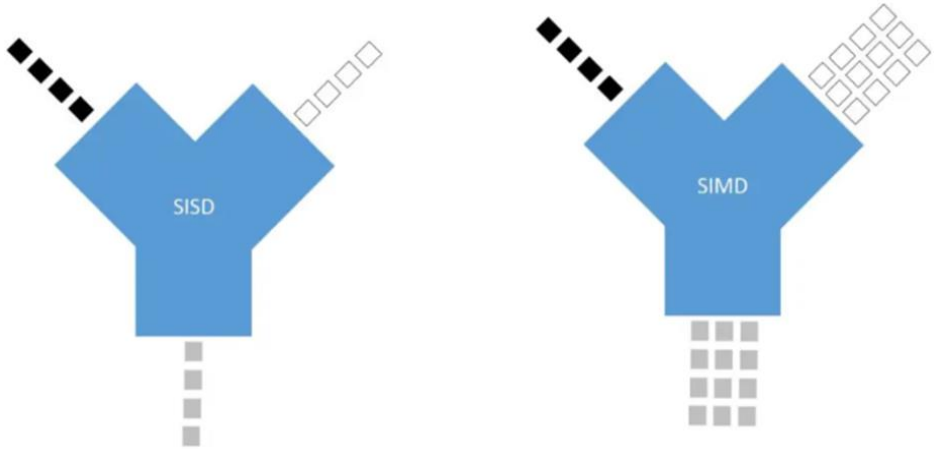
2022年基于RISC-V的SoC总市场规模达到16.3亿美元，预计2030年达到**920亿美元**，其中：

AI加速芯片将会成为所有RISC-V SoC中占比最高的部分，预计2030年总市场规模达到**420亿美元**，CAGR高达**49.2%**，是RISC-V增长预期中的重要组成部分



Source: The SHD Group, January 2024

RISC-V Vector指令标准概述



- SIMD相对SISD而言，单指令多数据
- 由图像，信号处理等应用催生
- 一般通过加大寄存器的位宽来应对并行数据

- 目前RISC-V定义了Vector扩展
  - V扩展，有单独的V寄存器和Ld/St指令，并行计算长度可变，适合高并行度计算场景

RISC-V Vector1.0 标准

芯来科技 NUCLEI

- 矢量扩展被称之为RV 指令集标准最重要的一组扩展，2015年发起，2021年正式生成标准。
- RVV 1.0 支持的数据类型广泛，运算类型丰富且可动态扩展，同一套指令可无修改适配各种微架构实现。
- RISCV GCC 从10.2 版本已经支持RVV1.0 指令，目前GCC 13 对应的intrinsic API 接口已经升级到最新 v 0.12 版本，且已部分支持自动向量化；预计GCC14正式发布后，GCC的自动向量化应该会完备。
- RISCV CLANG 17 版本也已支持最新 v 0.12 版本intrinsic API, 支持自动向量化。
- RISCV Linux 5.18 版本开始支持RVV，其它各种计算库及应用中间件都快速支持了RVV1.0。
- 有了RVV1.0 标准和软件生态的完备，为应对AI 对算力的需求，需要RISCV CPU 在微架构设计上做更多有针对性的设计。

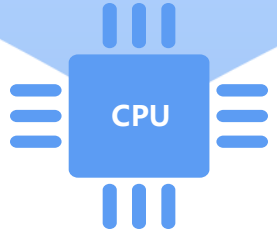
RISC-V CPU对于AI芯片的作用

芯来科技
NUCLEI

开发和调试　　安全和隐私

控制和调度　　数据预处理

通用计算能力



RISC-V®

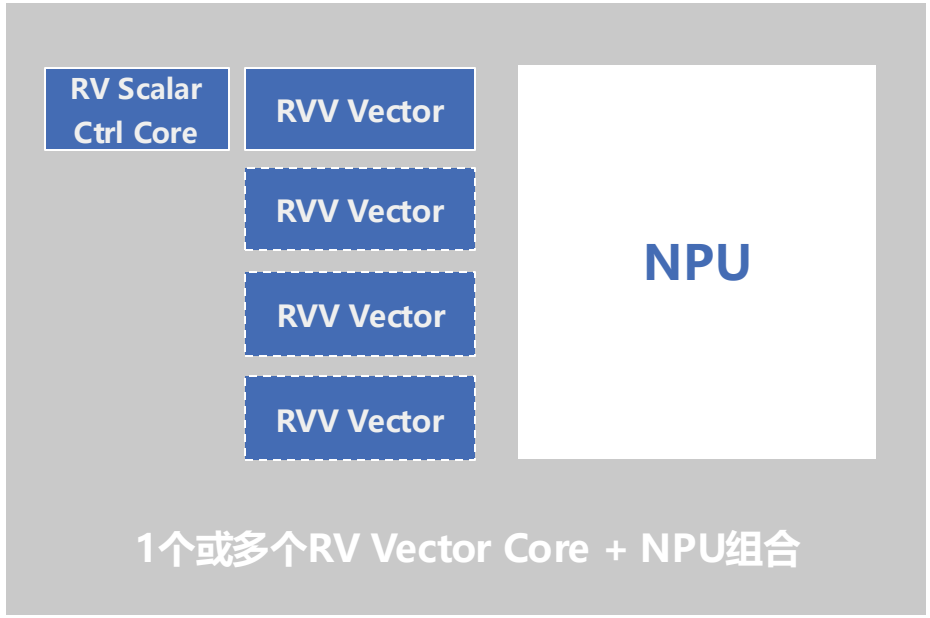
- **全球合作与标准化**：RISC-V的开放性促进了全球范围内的合作与标准化。不同国家和地区的公司和机构可以共同参与RISC-V的发展，推动AI芯片技术的全球标准化，促进互操作性和兼容性
- **丰富的可定制化**：不同的AI应用可能需要不同的处理能力和特性，公司和开发者可以根据自己的需求自由定制和优化处理器设计，通过高度可配置的RISC-V处理器，开发者可以优化AI算法的执行效率
- **生态系统支持**：随着RISC-V生态系统的不断壮大，越来越多的工具链、软件库和硬件设计支持RISC-V，为AI芯片开发提供了丰富的资源和支持，加快了开发和部署的速度
- **边缘计算的推动**：RISC-V的高效和可扩展性使其非常适合边缘计算设备。在边缘计算场景中，设备需要在功耗和性能之间找到平衡。通过使用RISC-V，AI芯片可以在保持低功耗的同时提供强大的计算能力，满足实时处理和低延迟的需求。

RISC-V可以加持赋能AI芯片中的CPU，更好的支持AI芯片的软硬件协同及生态

RISC-V已在AI芯片领域广泛落地

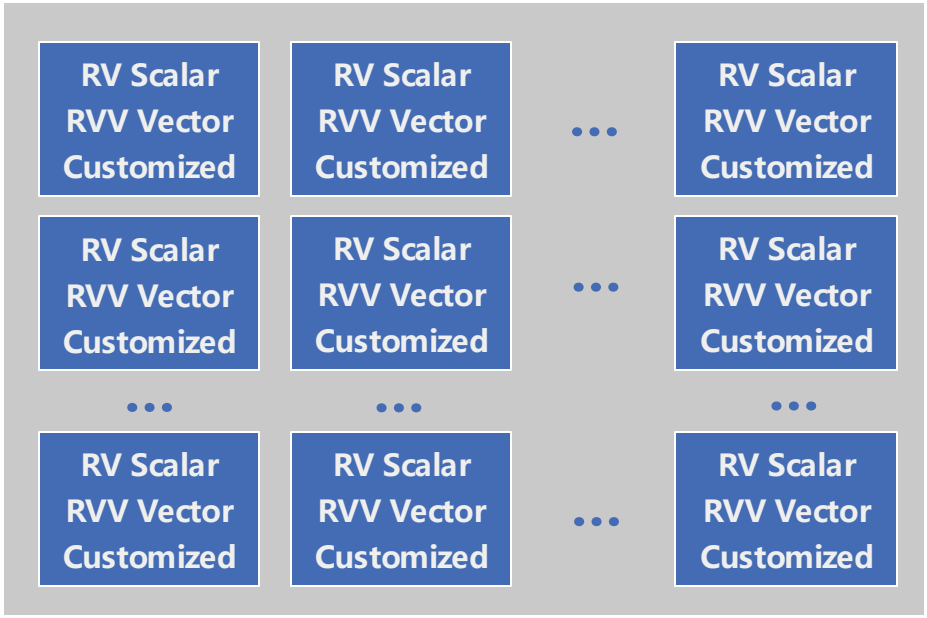
芯来科技 NUCLEI

典型方案一：RISC-V + NPU 方案



- NPU提供专用计算能力
- RVV Vector Core提供通用并行计算能力

典型方案二：定制RISC-V众核方案



- 基于RISC-V的RVV+自定义特定
- 众多的内核提供并行计算能力

AI芯片对于高并行度Vector Core的需求几乎是刚需
RISC-V RVV指令集的开放性与标准性，获得了业界广大支持以及丰富活跃的软件生态
基于RISC-V RVV的Vector Core可以很好的帮助NPU/GPU等加速器，提供更通用的性能

芯来科技 —— 本土专业的中立IP/解决方案提供商

芯来科技 NUCLEI

	北美	中国本土	其他
芯片/互联网巨头	intel NVIDIA MICROCHIP Meta Qualcomm	Alibaba 字节跳动 ZTE HUAWEI	NXP RENESAS SAMSUNG MEDIATEK
中立IP/解决方案提供商	SiFive MIPS SYNOPSYS 新思	芯来科技 NUCLEI	codasip ANDES TECHNOLOGY
芯片创业公司	VENTANA MICRO tenstorrent	RiVAI WITINMEM LeapFive ESPRESSIF ESWIN Lenxin Chipintelli 知合计算 TIMESINTELLI SUNLINE SPACEMIT HAAWKING ILSemi bluetrum VSILICO SOPHGO XCHIPET WCH沁恒	GREENWAVES TECHNOLOGIES Esperanto TECHNOLOGIES
开源提供商	Berkeley UNIVERSITY OF CALIFORNIA Western Digital	ICT 北京开源芯片研究院 BEIJING INSTITUTE OF OPEN SOURCE CHIP	PULP Platform
软件工具链	imperas ASHLING	ISCAS 卡姆派乐 COMPILER RT-Thread PerfXLab TeraPines 兆松科技	SEGGER LAUTERBACH Green Hills SOFTWARE IAR SYSTEMS

芯来科技简介

芯来科技
NUCLEI

- 芯来科技成立于2018年，是中国大陆本土专业的**RISC-V IP、子系统IP**及**SoC解决方案**提供商，赋能下游各类应用场景

5G通信　工业控制
AI　汽车电子
物联网　存储
MCU　网络安全

- 总部位于**上海，**在**华北、华东、华中**及**华南**均能够提供本地技术支持

- 从0到1自主研发**全系列RISC-V CPU IP，**拥有一流的特性和优势

丰富的可配置性
稳健的编码格式
完整的信息安全方案
良好的扩展性
健全的功能安全方案
灵活的商业模式及技术支持

- **累计出货量达数亿颗,**是国内本土RISC-V IP领军企业

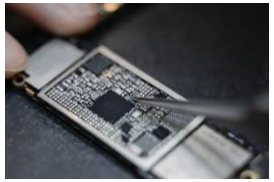
>220家正式授权客户

芯来科技已构建了全球最完备的RISC-V CPU IP产品货架之一

芯来科技 NUCLEI

通用处理器产品线

N 级别
32位架构
MCU, AIoT, 安全



U 级别
32位架构+MMU
Linux, 边缘计算

NX 级别
64位架构
存储, AR/VR

UX 级别
64位架构+MMU
Linux, 数据中心, 网络

专用处理器产品线

NS 级别
高安全性场景, 金融支付
SIM卡，物联网安全

NA 级别
ISO26262功能安全
汽车电子

NI 级别
人工智能，自动驾驶
通信计算，视频处理

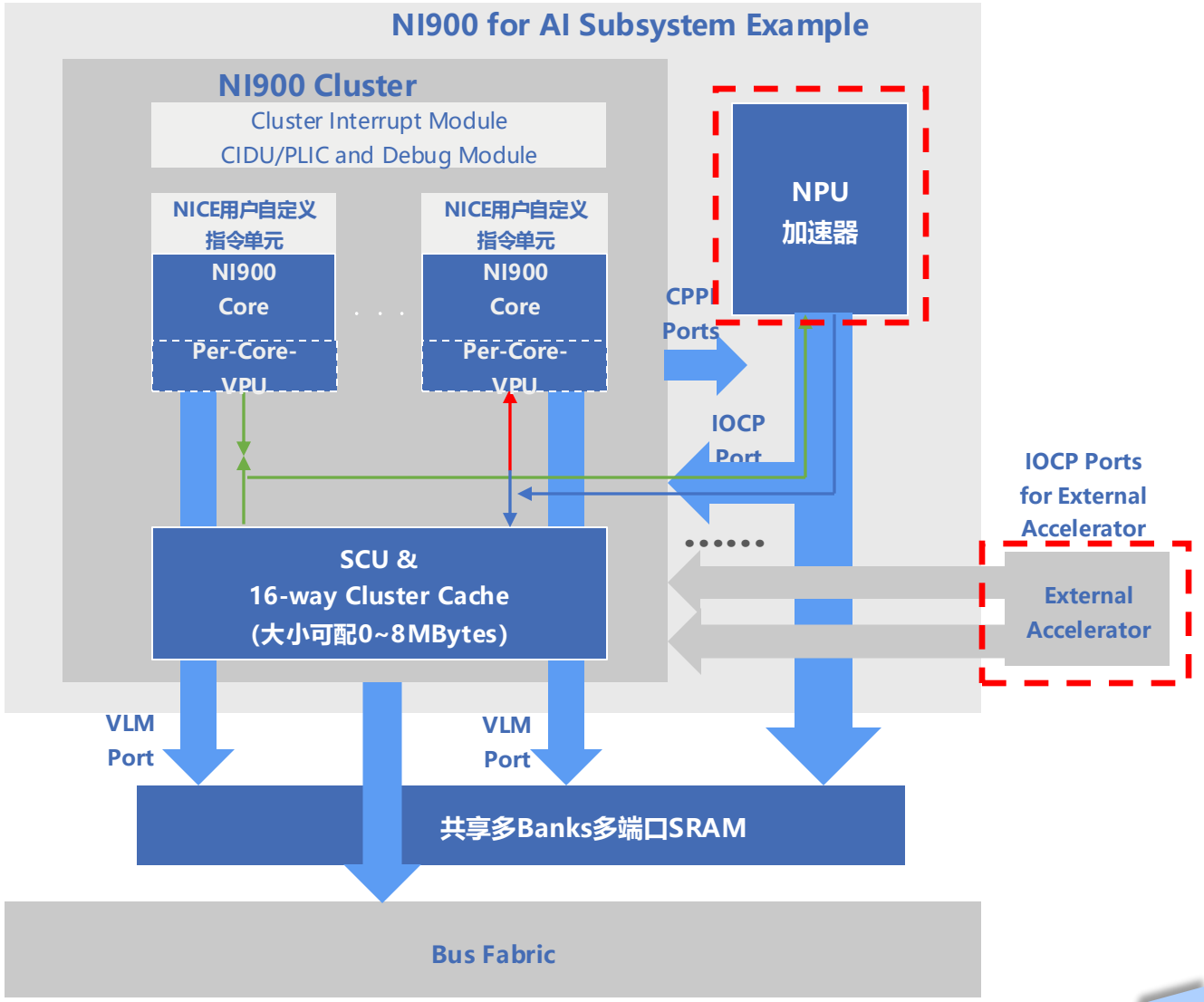
	N 级别	U 级别	NX 级别	UX 级别	NS 级别	NA 级别	NI 级别
1000 系列 Out-of-Order 3/4/6-Wide Decode				UX1000 (SMP)			
900 系列 9-Stage Pipeline Dual-Issue	N900 (SMP)	U900 (SMP)	NX900 (SMP)	UX900 (SMP)		NA900	NI900
600 系列 6-Stage Pipeline Single-Issue	N600	U600	NX600	UX600	NS600		
300 系列 3-Stage Pipeline Single/Dual-Issue	N300				NS300	NA300	
200 系列 2-Stage Pipeline Single-Issue	N200						
100 系列 2-Stage Pipeline Single-Issue	N100				NS100		

芯来科技NI900，为国产AI芯片赋能

芯来科技 NUCLEI

NI900并非取代NPU、GPU等加速器，而是与之进行结合，更好的服务SoC上的加速器，提供AI芯片公司更完善的整体解决方案

- **基础标量处理器：**可以配置为900系列的RV32或RV64的任何一款——N900/U900/NX900/UX900
- **RVV1.0 VPU：**可配置基于RISC-V V Extension（RVV1.0 Vector指令集）的VPU单元，VPU的VLEN可配置为512b或者1024b
  - NI900支持Cluster内的每个Core均可以配置VPU
- **NPU加速器：**可通过NI900的IOCP（IO Coherent Port）与处理器紧耦合，实现对CPU内部Cache的一致性
- **用户自定义指令扩展接口：**用户可以使用Nuclei的NICE硬件扩展接口，增加自己自定义的指令，包括Scalar或Vector指令



芯来科技AI DSA NI900处理器

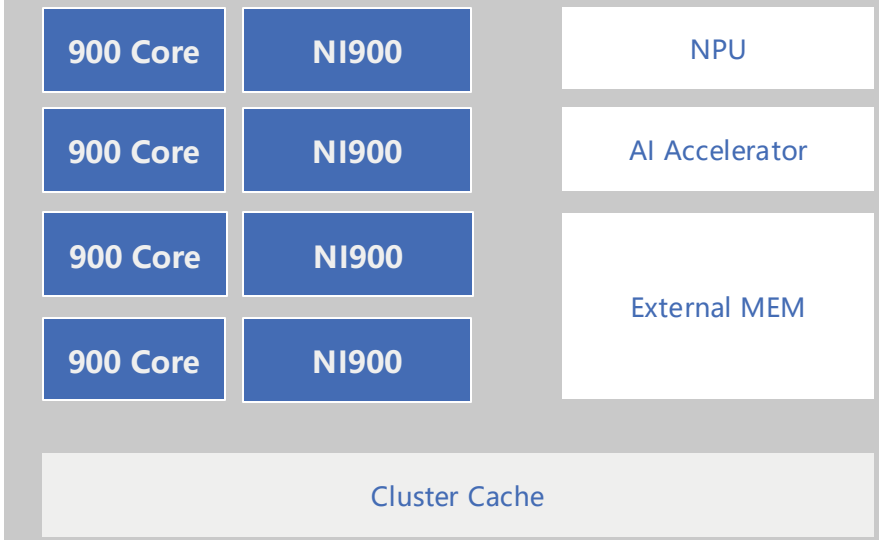
芯来科技
NUCLEI

针对AI应用

安防摄像头

推理

数据中心AI加速



	NI900处理器特性
RVV 1.0	VLEN = 512b & 1024b
EXE Width	Max 4 instructions executing (scalar + vector)
NICE	Both scalar and vector customized instruction supported
Data Type	INT8/16/32, BFP16/FP16/FP32
VPU Configuration	Configurable per-core, partial and shared VPU
Vector Mem-subsystem	VLM port for excellent memory access performance
External Devices	Support NPU, AI accelerator and other external devices deeply coupled with main CPU
Data Subsystem	Data Cache, DLM , Private L2 Cache, Cluster Cache
SoC Connectivity	IOCP port for SoC level coherent access
Dual Mode	Application + Real-Time

NI900的VPU可进行多种参数配置

芯来科技
NUCLEI

参数描述

- VLEN：一个向量寄存器的总bit数（宽度）
- DLEN：运算单元能够并行处理的一个向量元素的最大bit数
- ELEN：并行处理的数据类型的最大宽度，如果ELEN=32，则最大的处理数据类型是INT32和FP32

可配选项	参数值
VLEN_512	VLEN = 512, DLEN = 512, ELEN = 32/64
VLEN_1024	VLEN = 1024, DLEN = 1024 ELEN = 32/64

NICE（Nuclei Instruction Co-unit Extension)

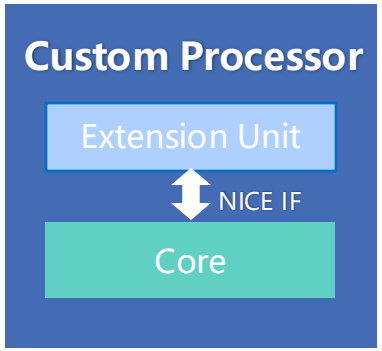
芯来科技 NUCLEI

NICE机制使得用户可以结合自己的应用扩展自定义指令，将芯来的标准处理器核扩展成为面向领域专用的处理器。芯来科技的NICE机制不仅可以进行运算型的自定义指令扩展，且自定义指令模块还可以访问Core的存储资源（DCache等）实现与主Core的内存一致性

1. Define user-define instructions

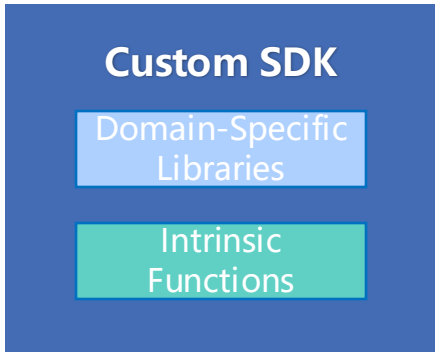
- Identify user-defined instructions by program profiling
- Define extension instructions into RISC-V reserved ISA space

2. Implement Extension Unit



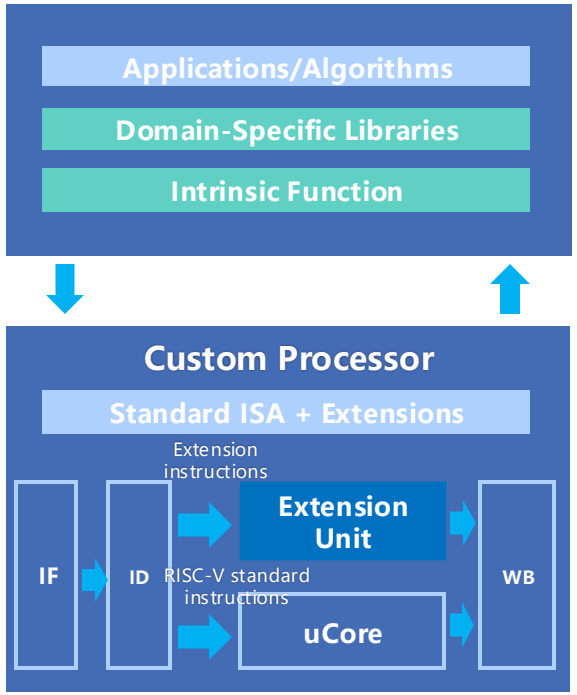
- Implement the application-specific co-unit following the NICE interface

3. Develop Domain Specific Lib/Function



- No requirement for tool-chain update
- Using Intrinsic Function or encapsulated libraries

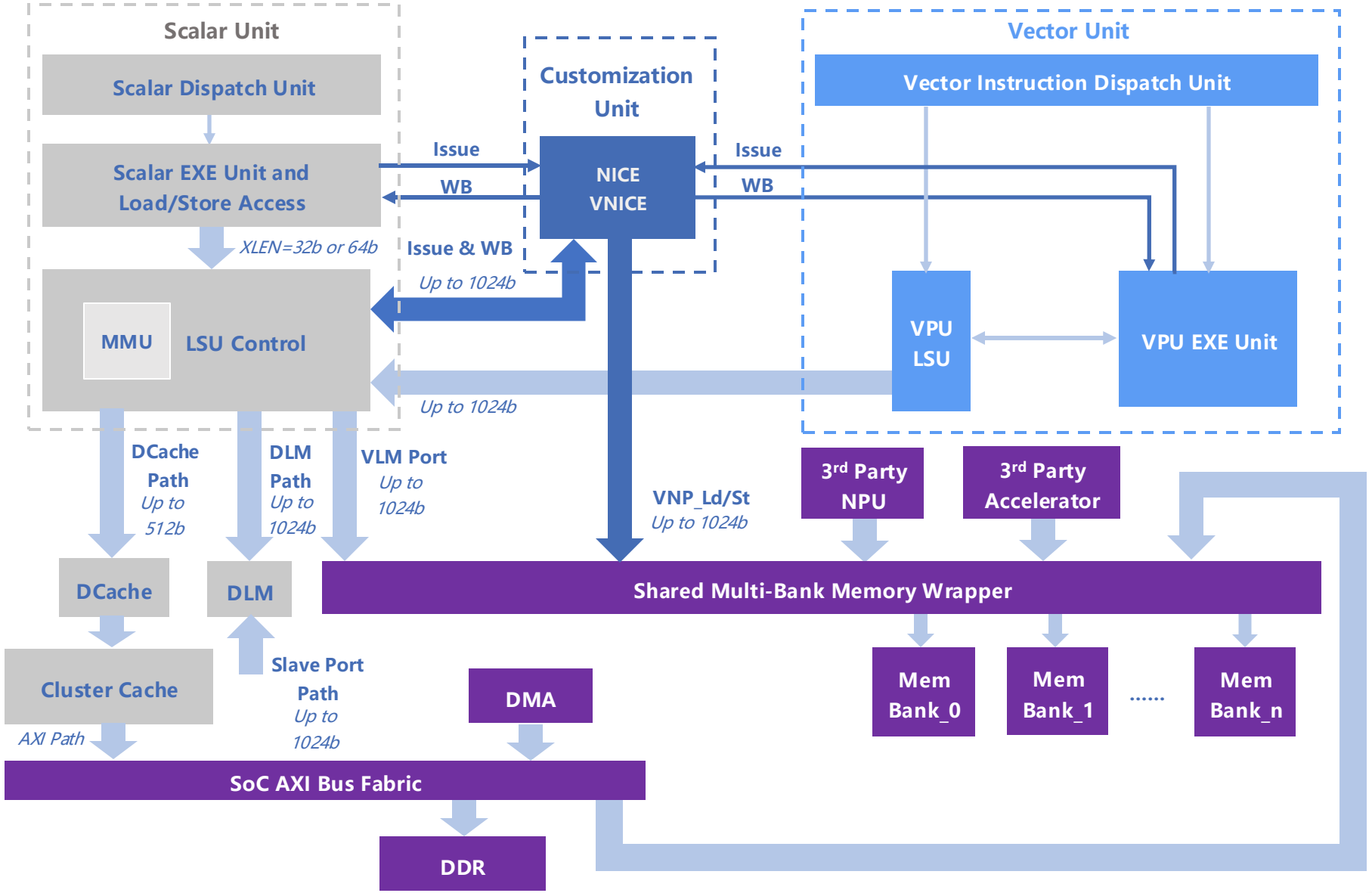
4. Accelerate Domain Specific Applications



NI900 VPU提供强大的向量访存解决方案

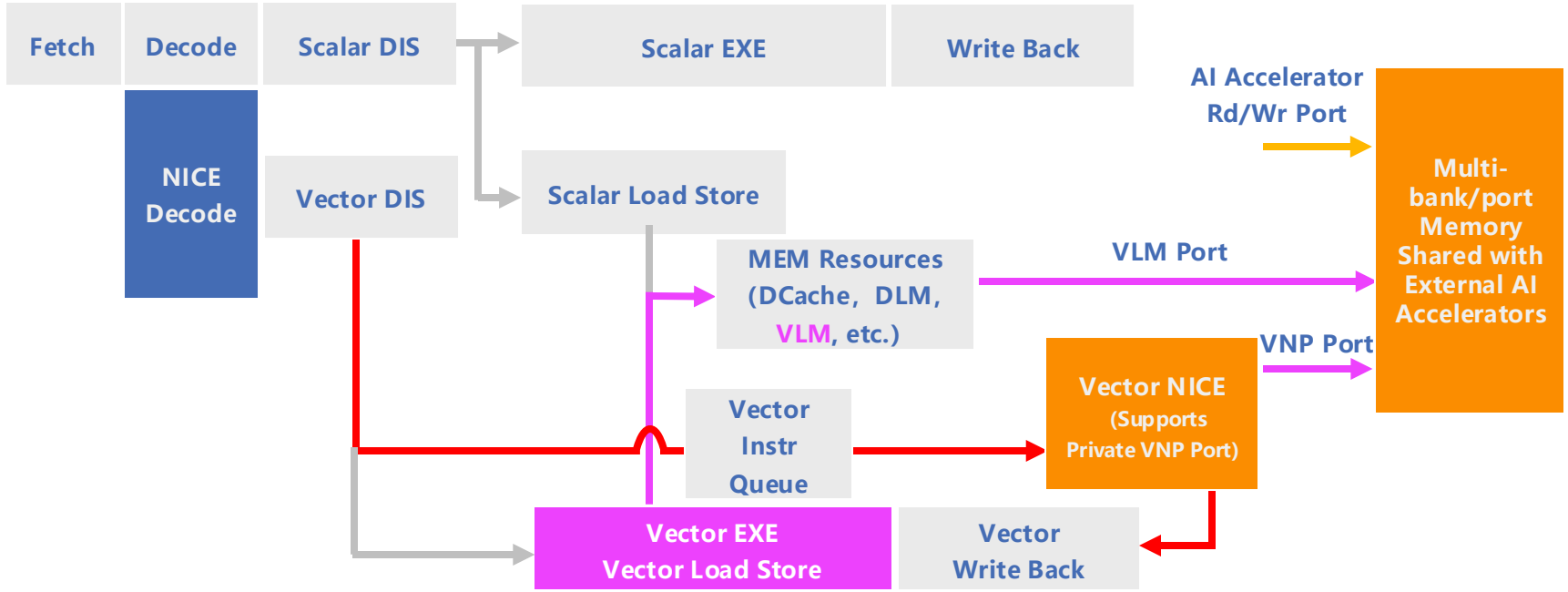
芯来科技 NUCLEI

- 芯来NI900拥有完整的**标量 + 向量 + 标/向量自定义指令接口**系统解决方案
- 标量、向量和自定义指令接口被**深入的整合在一起**
- 整个内存资源被完美的分配到全系统



NI900 VPU拥有极其灵活的可配置性

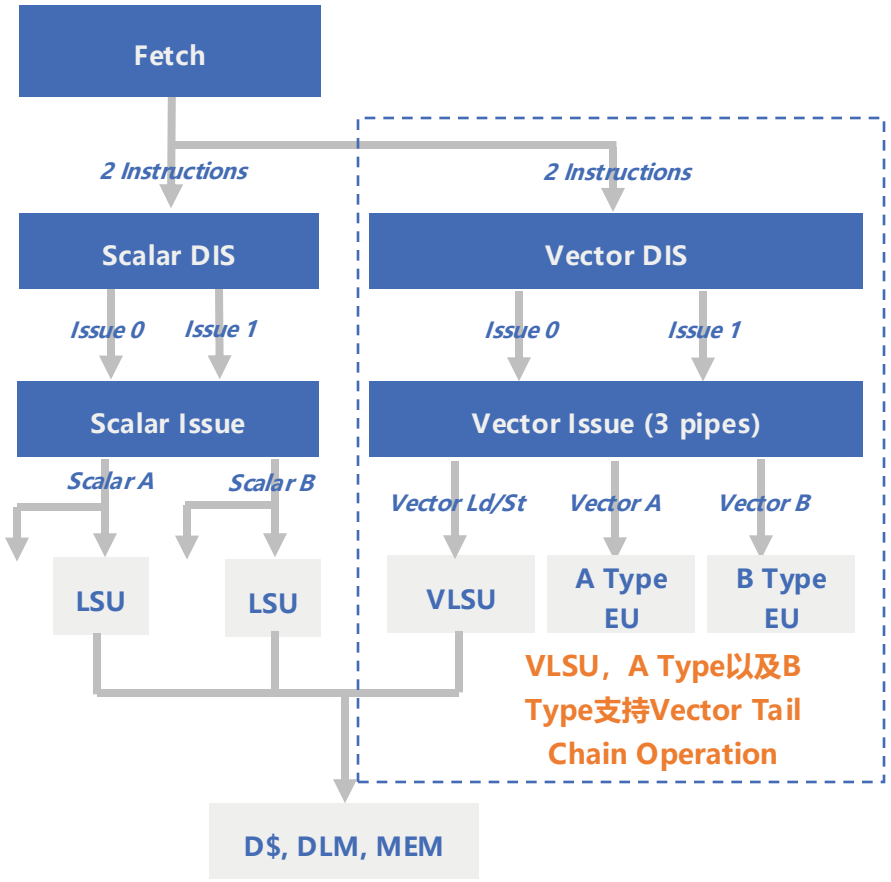
芯来科技
NUCLEI



- **Full VPU + VNICE**
  - 实现全RVV指令的VPU + 向量自定义指令单元，提供最强的向量计算能力
- **Customized VPU + VNICE**
  - 实现部分RVV指令的VPU + 向量自定义指令单元，通过裁剪掉部分向量指令来提供面积和功耗性价比更高的解决方案
- **VNICE only**
  - 向量Load/Store单元 + 向量自定义指令单元，只实现需要使用的向量指令

NI900 VPU提供优秀的并行计算性能

芯来科技
NUCLEI

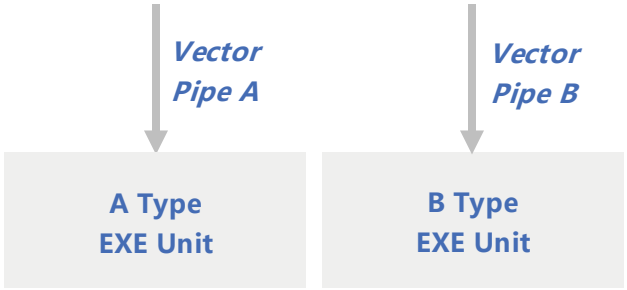


NI900 双发射机制：

- 标量指令可单独双发射
- 向量指令可单独双发射
- 单个标量和向量指令可以同时发射
- 标量和向量指令可以乱序执行
- 标量和向量单元共享内存资源 (DCache，DLM，External Memory, etc)

NI900 VPU提供优秀的并行计算性能

芯来科技 NUCLEI



NI900 VPU向量Tail Chain操作

RVV中, 类似LMUL等指令需要多个周期才能完成计算

Instruction A: Load data to V0~V3, takes 4 cycles to finish

Instruction B: Add V0~V3 with V4~V7

In this case, Inst B cannot start unless Inst A finishes

	Cycle	0	1	2	3	4	5	6	7
Without Tail Chaining	A	V0	V1	V2	V3				
	B					V0 + V4	V1 + V5	V2 + V6	V3 + V7
With Tail Chaining	A	V0	V1	V2	V3				
	B		V0 + V4	V1 + V5	V2 + V6	V3 + V7			

without vector tail chaining mechanism, pipe B must wait until all elements finish

Takes 8 cycles to finish Inst B

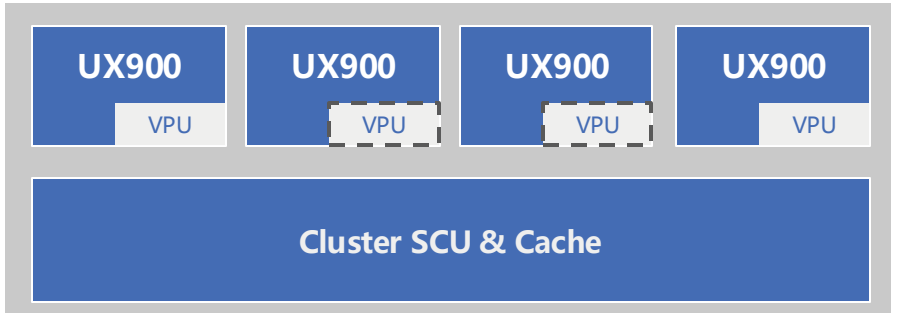
with vector tail chaining mechanism, each element can be forwarded from pipe A to pipe B

Takes 5 cycles to finish Inst B

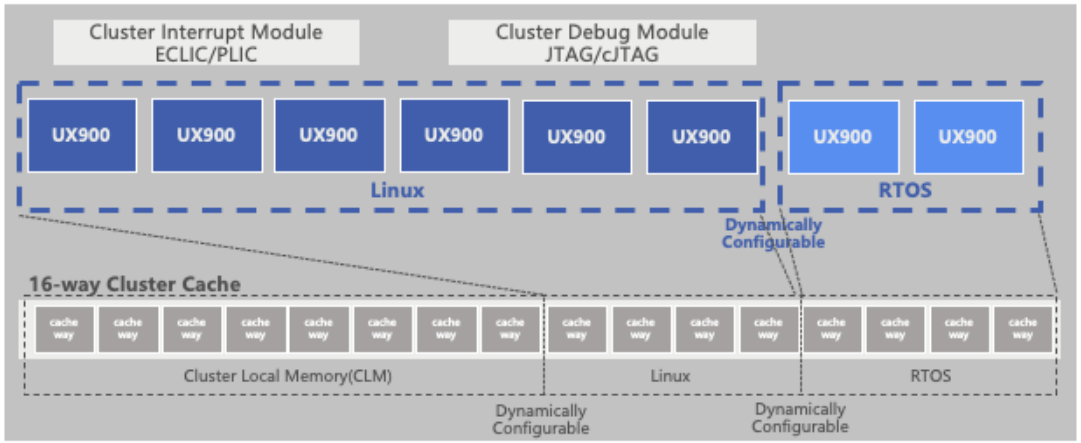
NI900提供多样化的Cluster级别解决方案

芯来科技 NUCLEI

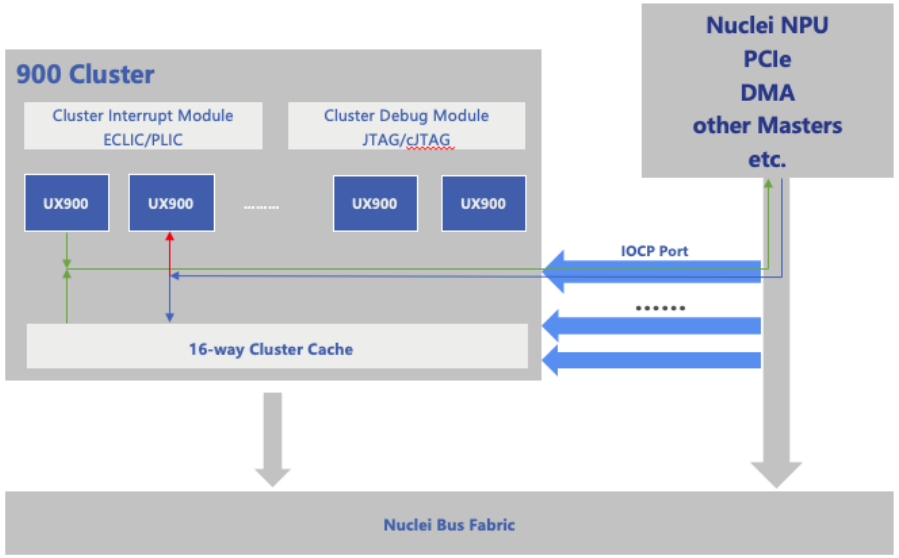
不同的VPU Cluster配置



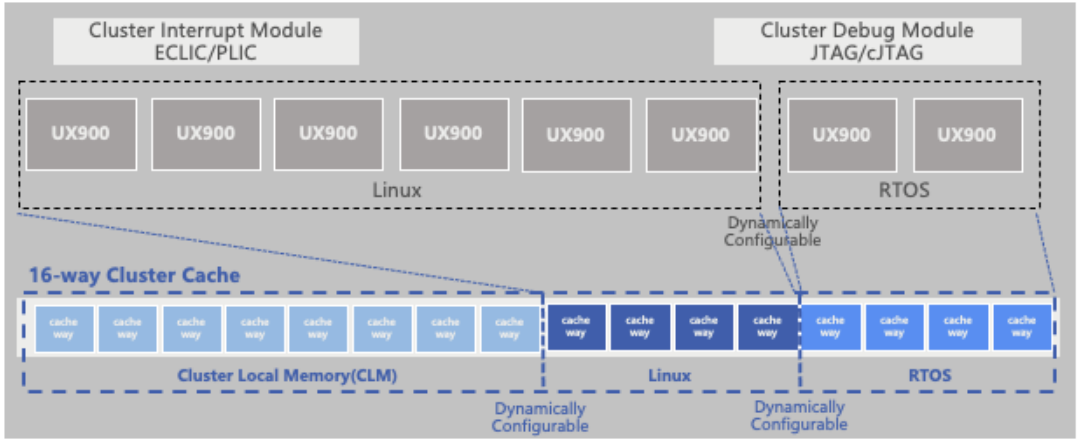
NI900双模式: 应用 + 实时



NI900 IOCP接口可以深度整合外部加速器



NI900 CLM模式



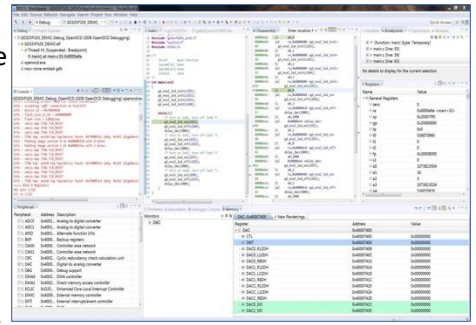
NI900提供高可适配的软件生态

芯来科技 NUCLEI

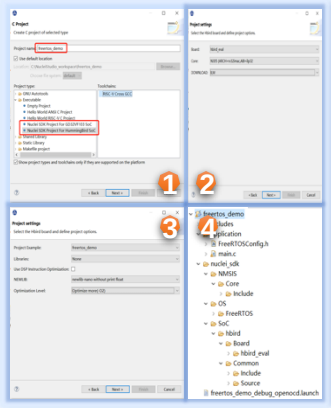
Nuclei Studio IDE

- Eclipse CDT based
- Integrated GCC and OpenOCD
- Nuclei SDK deeply integrated
- Libre and free
- Portable executables, without installation
- Easy-to-use project template
- Easy Project Configuration
- Integrated editor
- In system debugging
- In system programming
- Integrated serialport tool
- Real time register display
- Support Linux and Windows

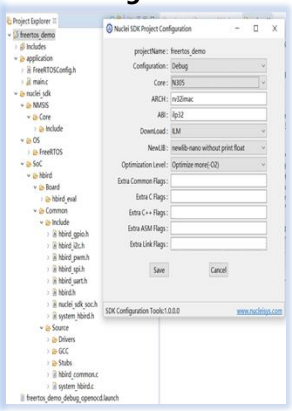
IDE Interface



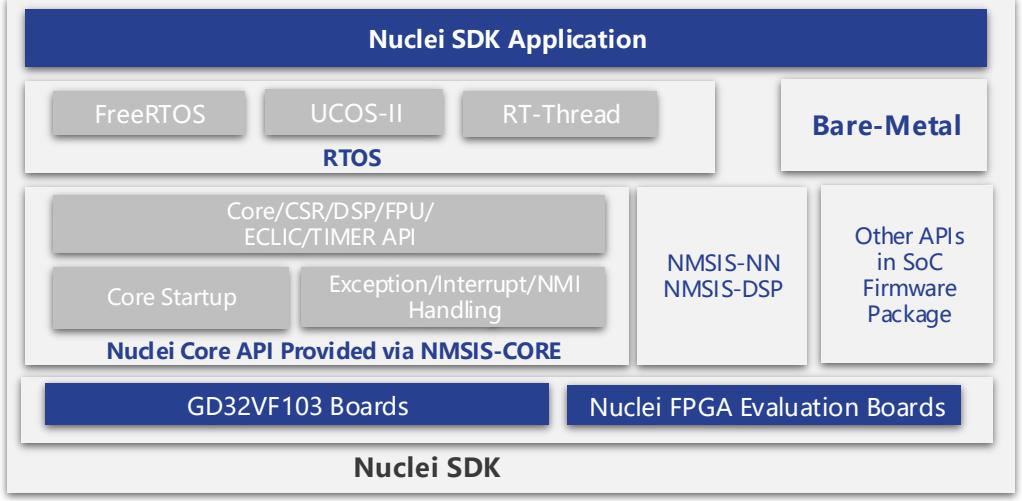
Nuclei SDK Project Template



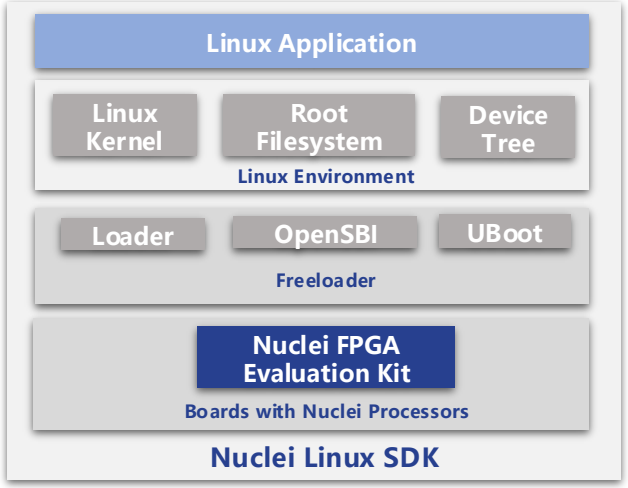
Nuclei SDK Project Configurations



芯来软件SDK



芯来Linux SDK



芯来调制器和开发板

Hummingbird

蜂鸟调试器

DLINK Debugger

DDR200T

MCU200T

RV-STAR

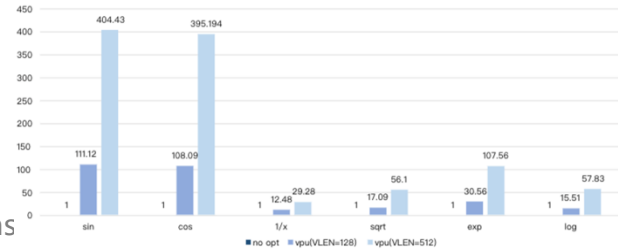
CM32M433R-START

Nuclei DSP, NN, Crypto软件库

Optimized DSP, Neural Network and Crypto Libraries specifically for Nuclei's CPUs

Nuclei Libraries includes:

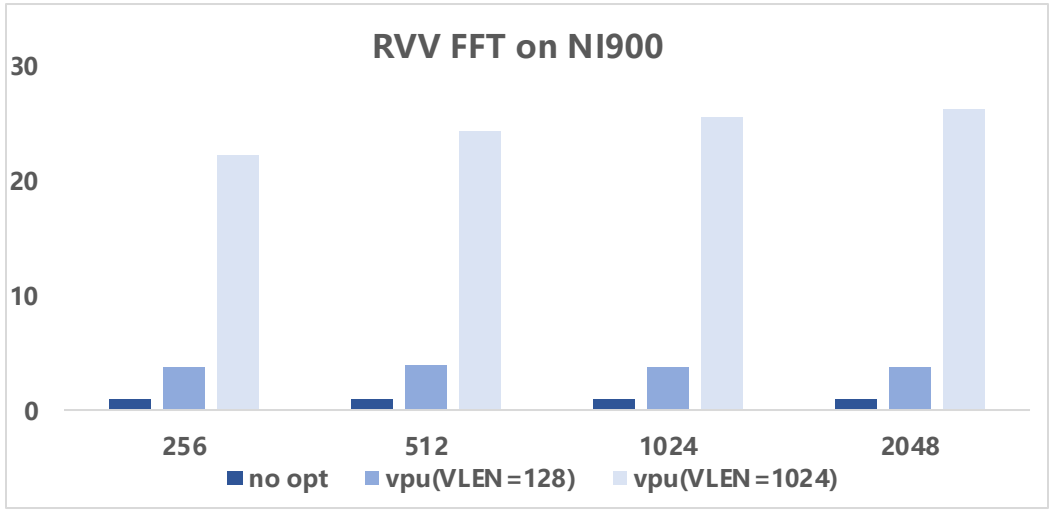
- AES algorithm
- DES
- Examples including
- Basic math functions
- Bayesian estimators
- Complex math functions
- Controller functions
- Fast math functions
- .....

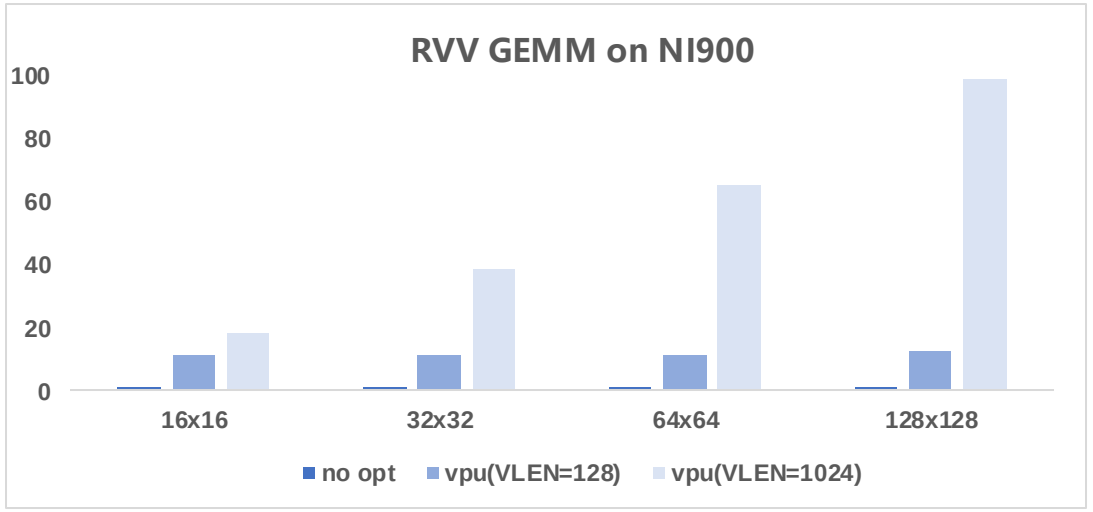


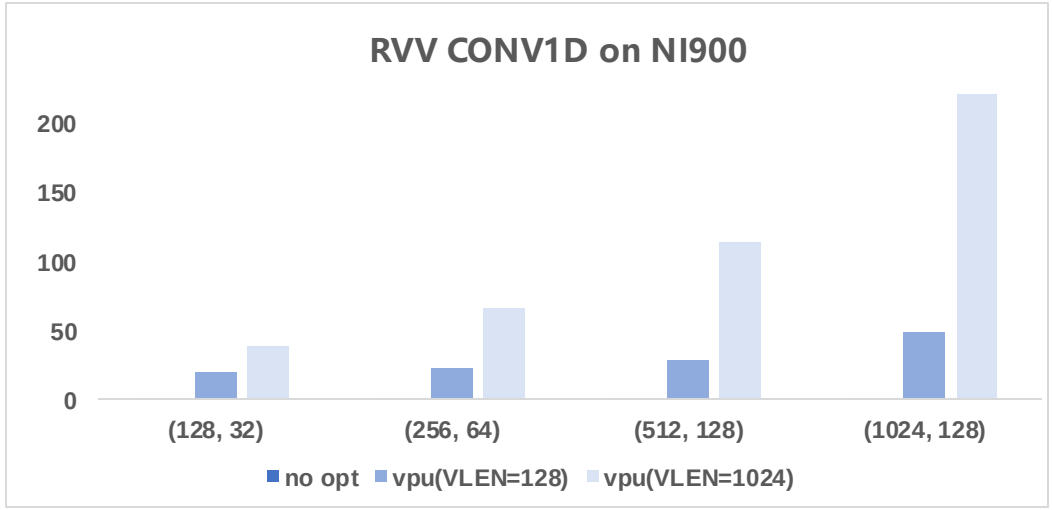
Huge Performance Enhancement after DSP Libraries Optimization

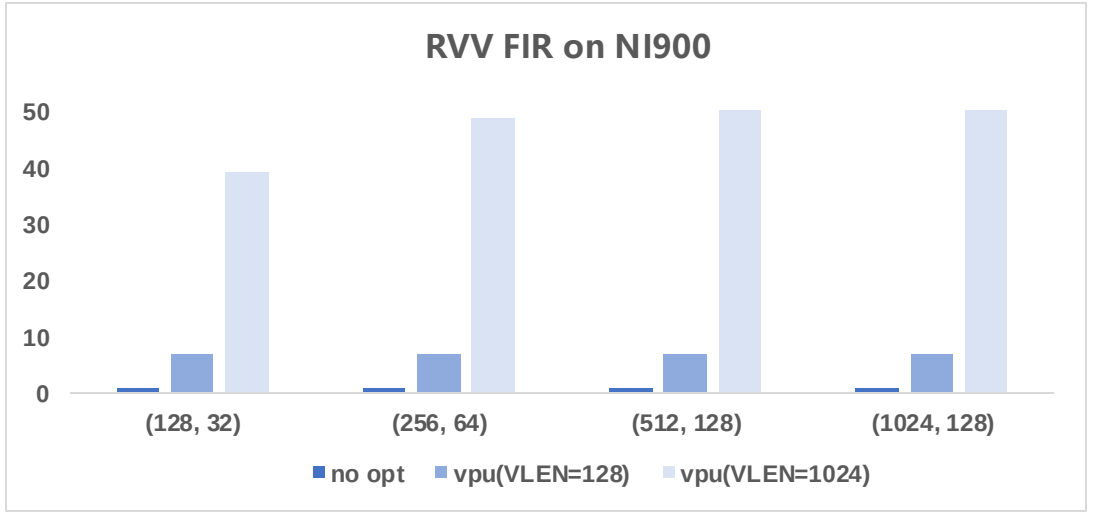
芯来科技NMSIS DSP库性能优化

芯来科技
NUCLEI



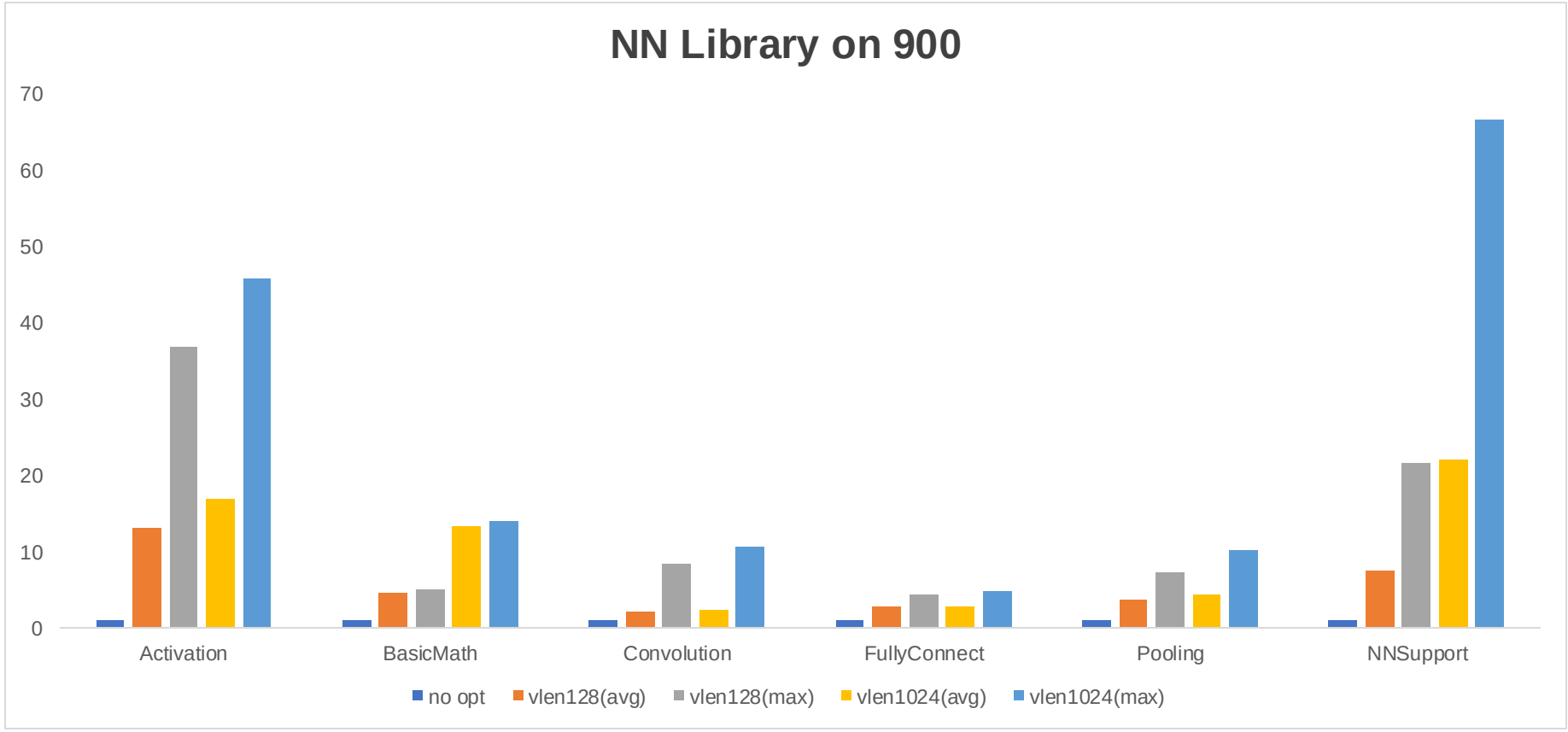






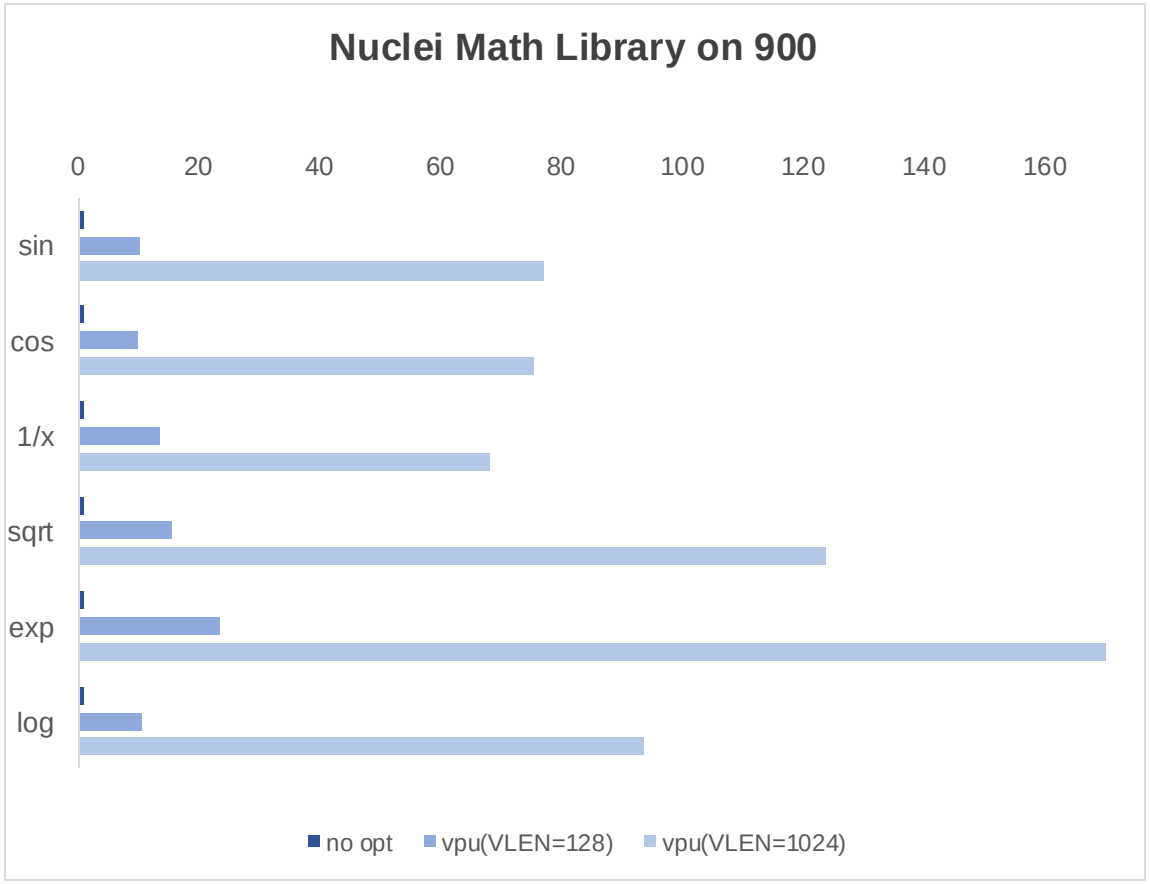
Accelerates FFT, GEMM, CONV1D, and FIR functions with RVV，x-axis：Data length, y-axis: Speedup

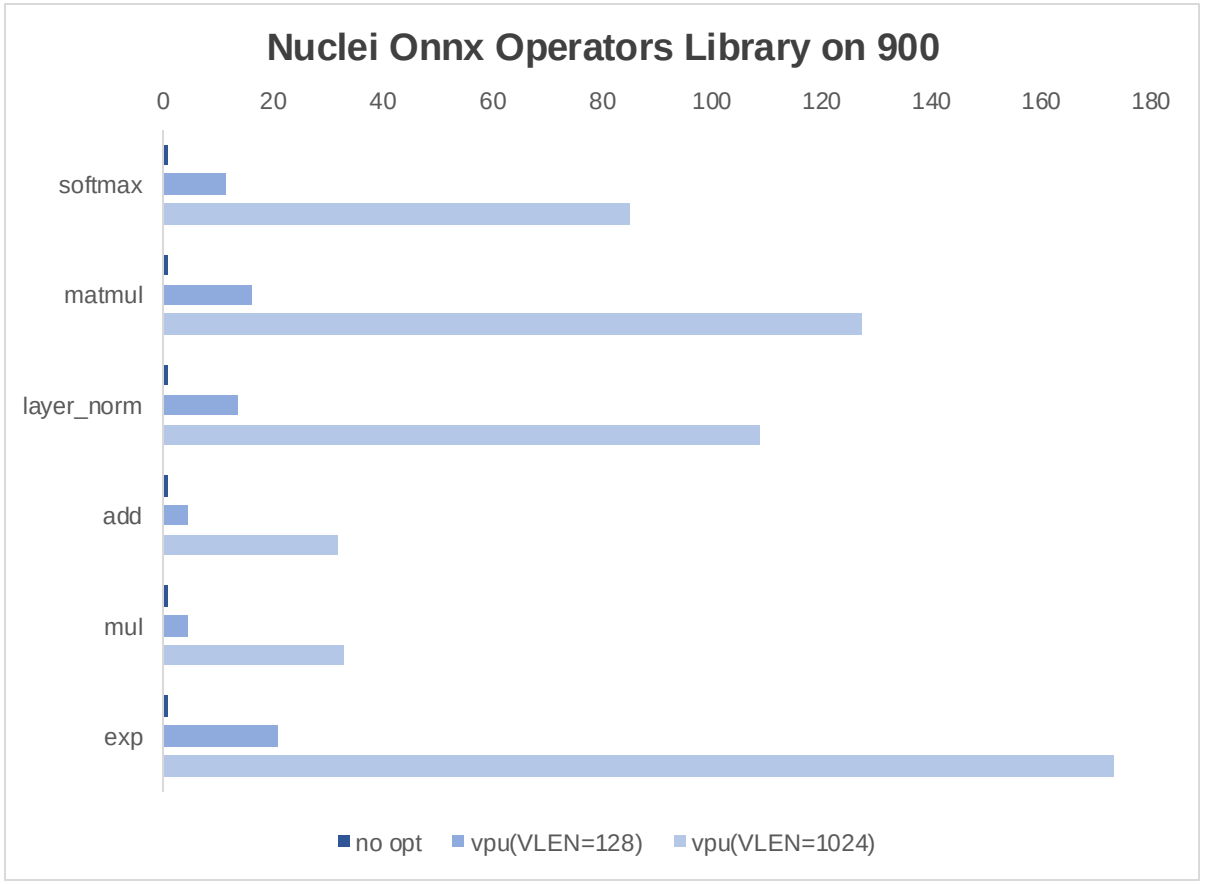
芯来NMSIS NN库性能优化



X-Axis: NMSIS DSP/NN Library Algorithm Type
Y-Axis: Performance Enhancement

芯来Math & Onnx操作库优化

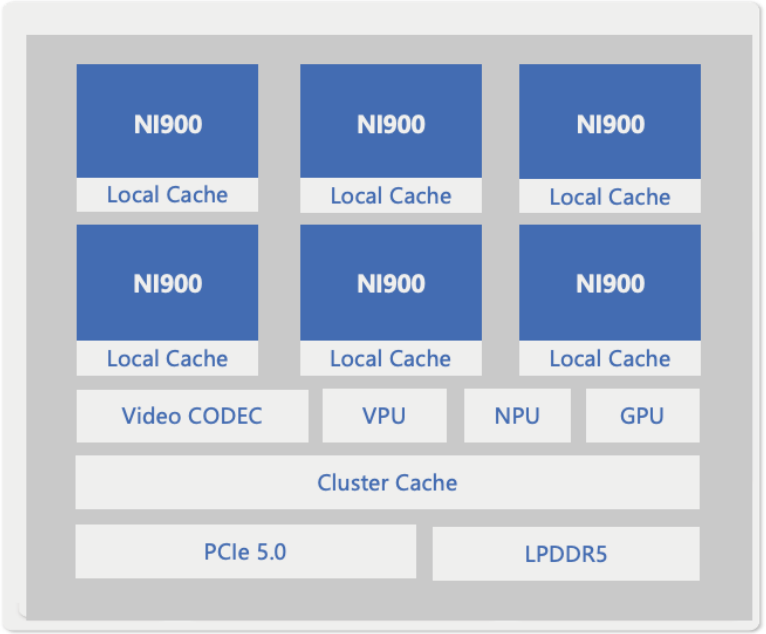




Accelerate math functions with RVV，y-axis: Speedup

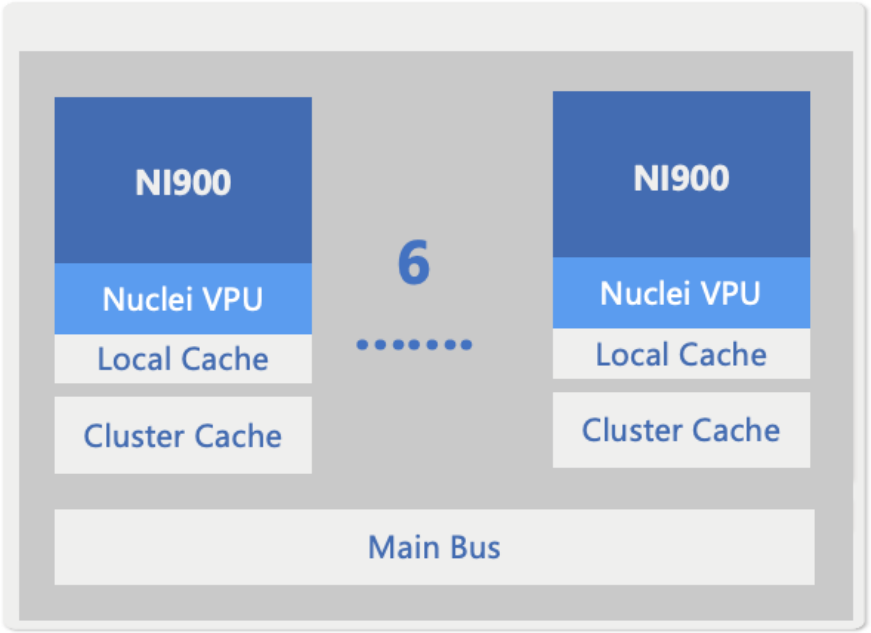
NI900部分客户案例

芯来科技
NUCLEI



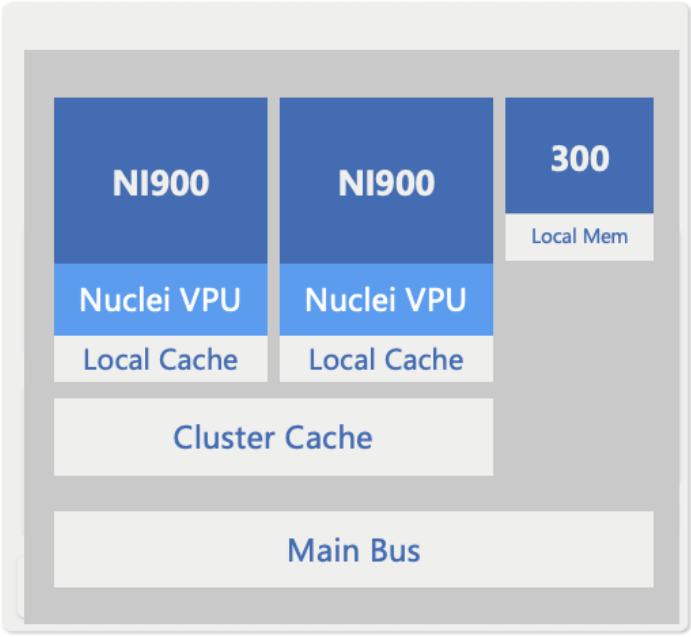
云端视频加速 SoC

6核 NI900



存算一体AI SoC

8核 NI900



AI存储芯片

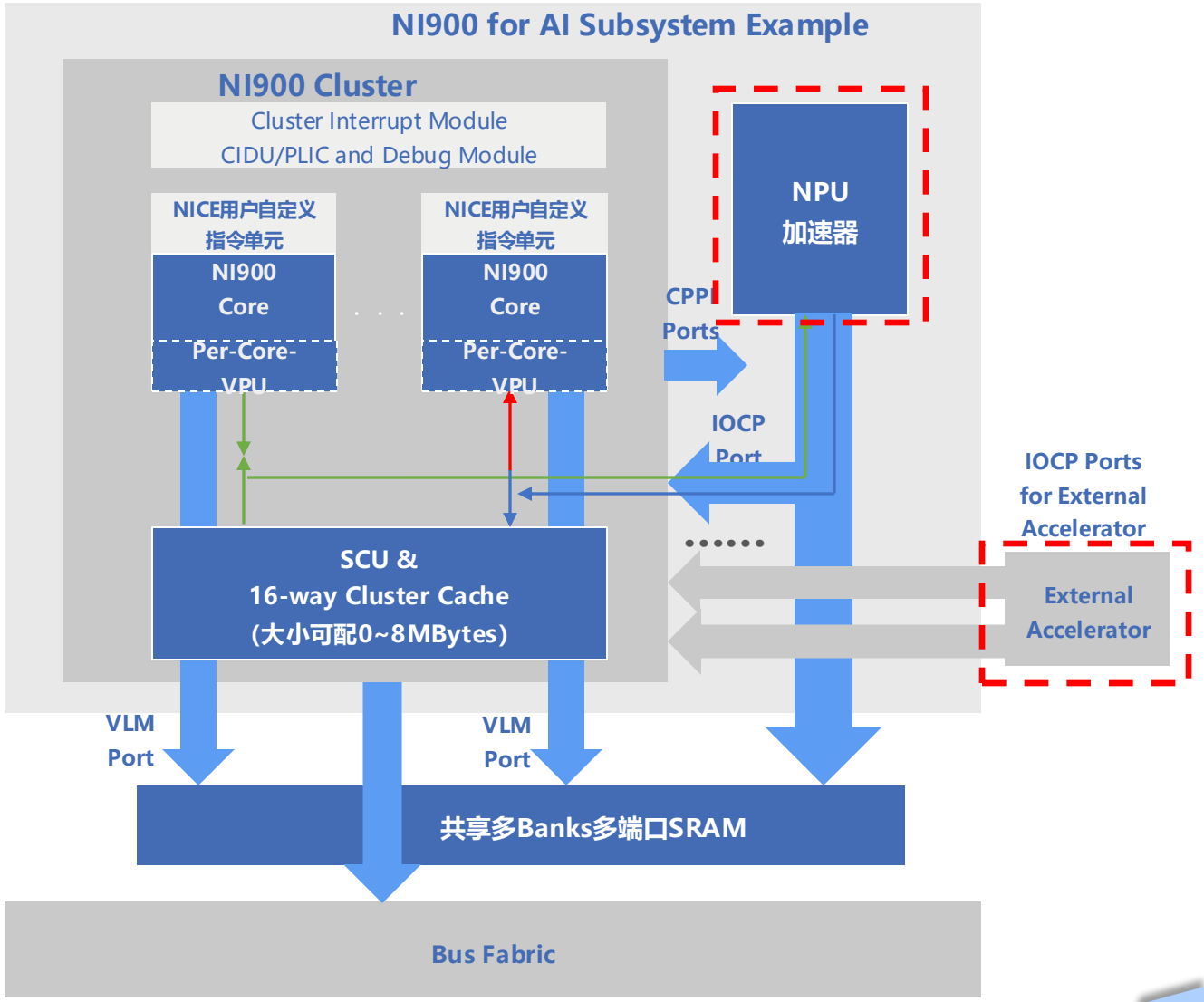
双核NI900 + N300 大小核

芯来科技NI900，为国产AI芯片赋能

芯来科技 NUCLEI

NI900并非取代NPU、GPU等加速器，而是与之进行结合，更好的服务SoC上的加速器，提供AI芯片公司更完善的整体解决方案

- **基础标量处理器：**可以配置为900系列的RV32或RV64的任何一款——N900/U900/NX900/UX900
- **RVV1.0 VPU：**可配置基于RISC-V V Extension（RVV1.0 Vector指令集）的VPU单元，VPU的VLEN可配置为512b或者1024b
  - NI900支持Cluster内的每个Core均可以配置VPU
- **NPU加速器：**可通过NI900的IOCP（IO Coherent Port）与处理器紧耦合，实现对CPU内部Cache的一致性
- **用户自定义指令扩展接口：**用户可以使用Nuclei的NICE硬件扩展接口，增加自己自定义的指令，包括Scalar或Vector指令



芯来科技
NUCLEI

芯来科技公众号

芯来科技业务联络

谢谢您！