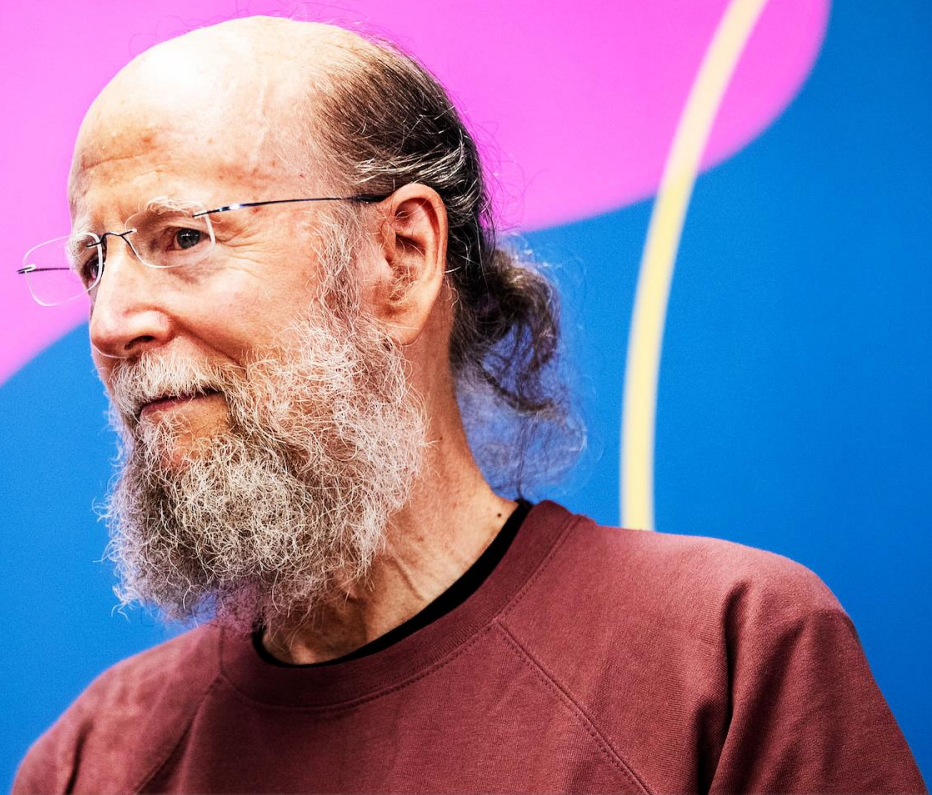
Imagination GPU and Risc-V Enable Industrialisation and Eco-Innovation

Speaker:
Zack.zheng@imgtec.com,
Director of Product
Imagination

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.”

从70年的人工智能研究中可以得出的最大教训是，利用计算的通用方法最终是最有效的，而且差距非常大。

Rich Sutton | The bitter lesson



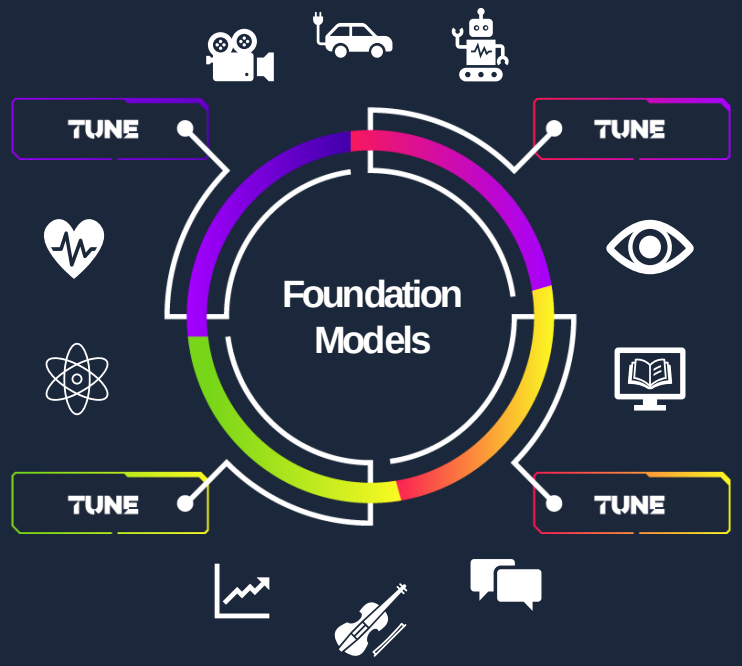
SUCCESS AT THE EDGE

AI playbook now requires efficiently scaling edge compute and at the same time driving de-fragmentation of architectures and solutions.

AI at the edge will have to be developed in accordance with the tenets of the Bitter lesson in order to leverage advances in foundation model capabilities in language, speech, video, autonomy, health etc.

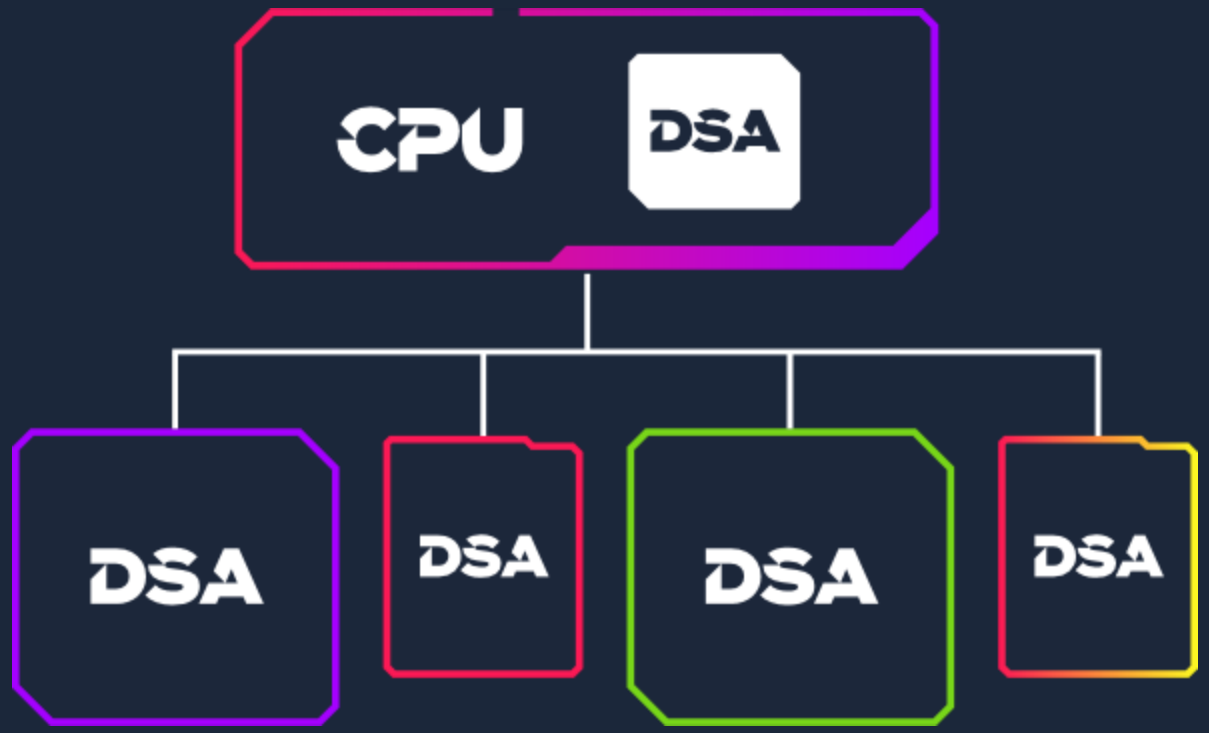
How should it be done?

最终通用计算方法：
基础模型 + 开放软件生态系统



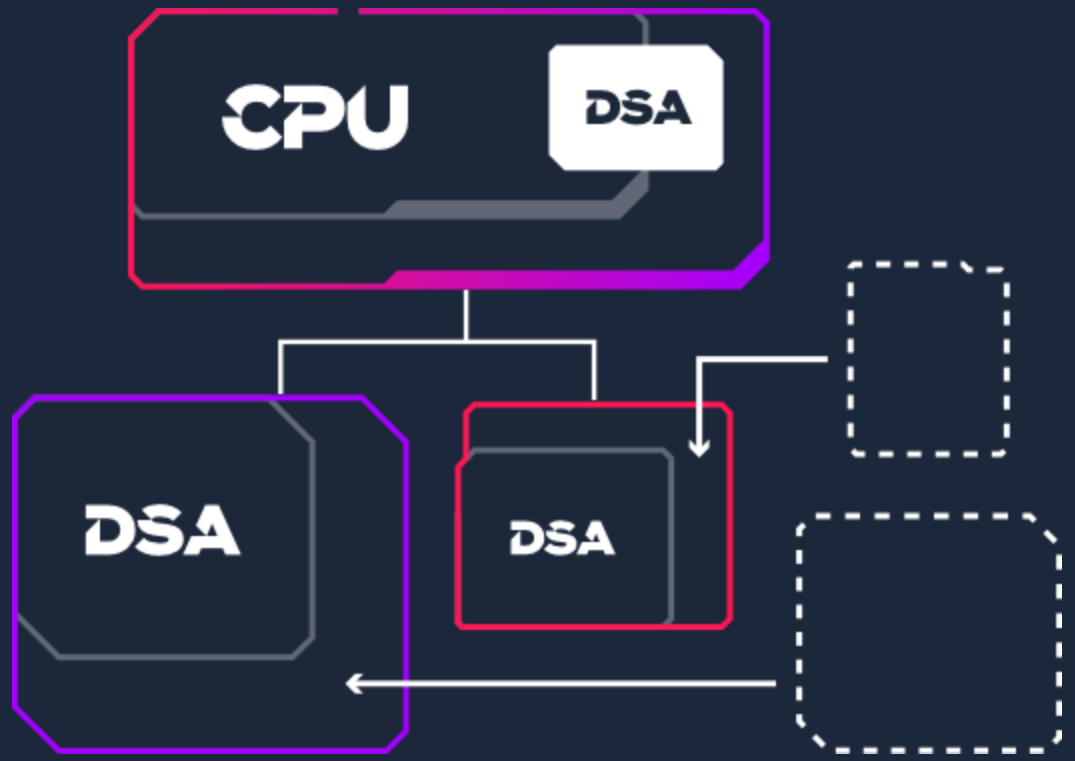
TensorFlow
SYCL
OpenCL
tvm
PyTorch
oneAPI
Vulkan

传统的端侧架构



Spend area on Domain Specific Accelerators under a dark silicon assumption to target acceleration-level KPIs
将资源投入到特定领域加速器上，以实现加速水平的关键绩效指标（KPI）

新的系统架构理念

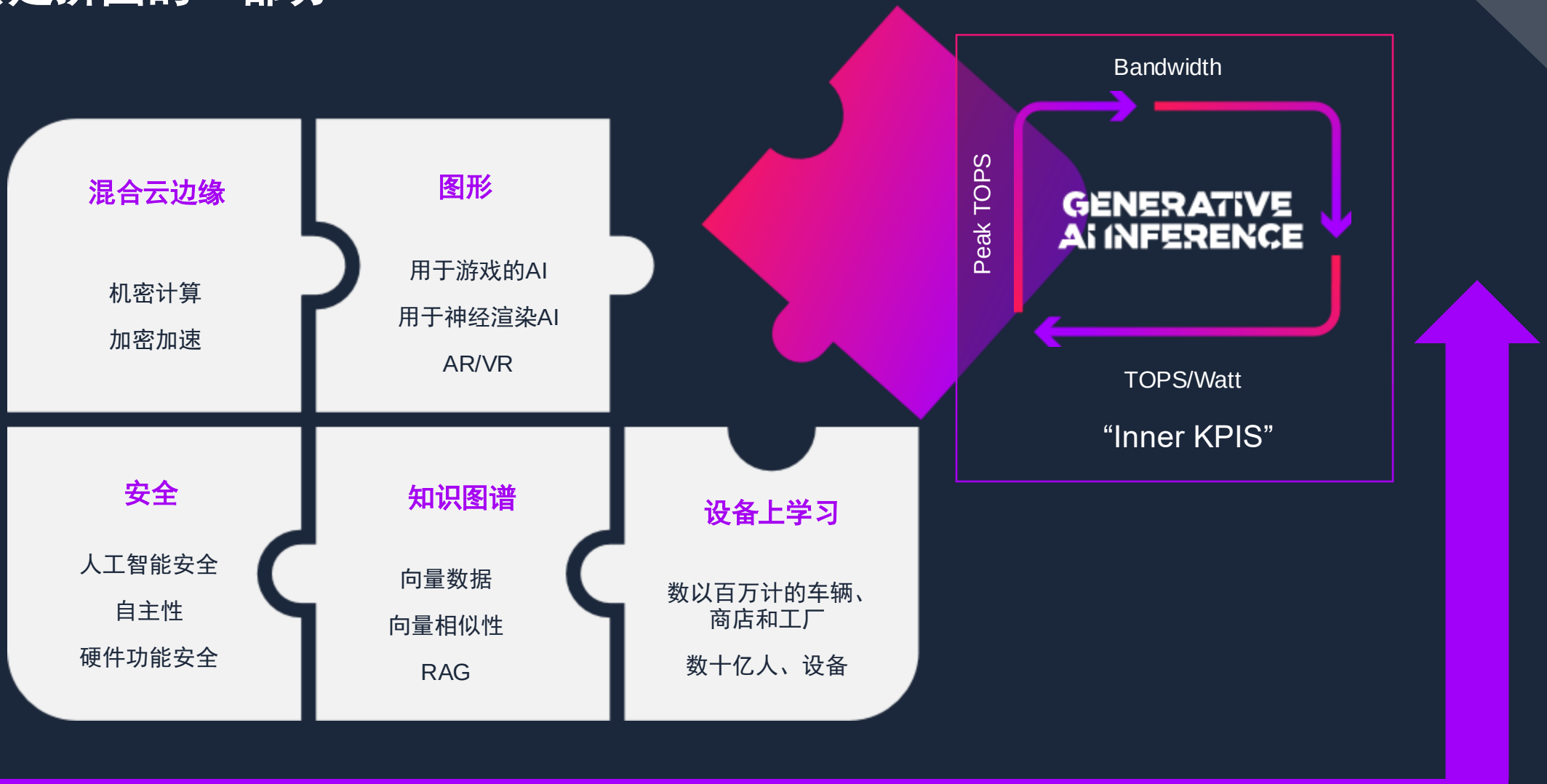


Spend area saved on multiple Domain Specific Accelerators on enhancing the energy-efficiency of most general-purpose accelerators and computing cores

将多个特定领域加速器节省下来的面积用于提高大多数通用加速器和计算内核的能效

对于边缘能源效率需要整体方法

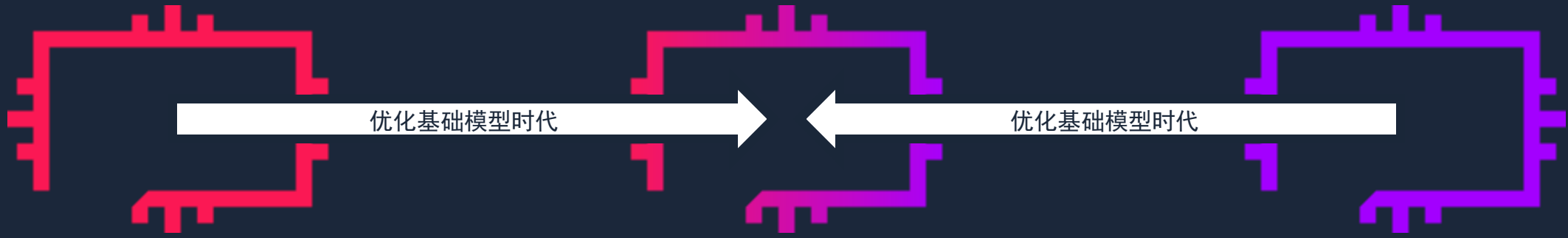
推理只是拼图的一部分



其中部分或全部需要有效加速，并与用户体验完美融合

边缘计算解决方案
优化基础模型

传统 GPU，CPU 起点



传统 NPU 起点

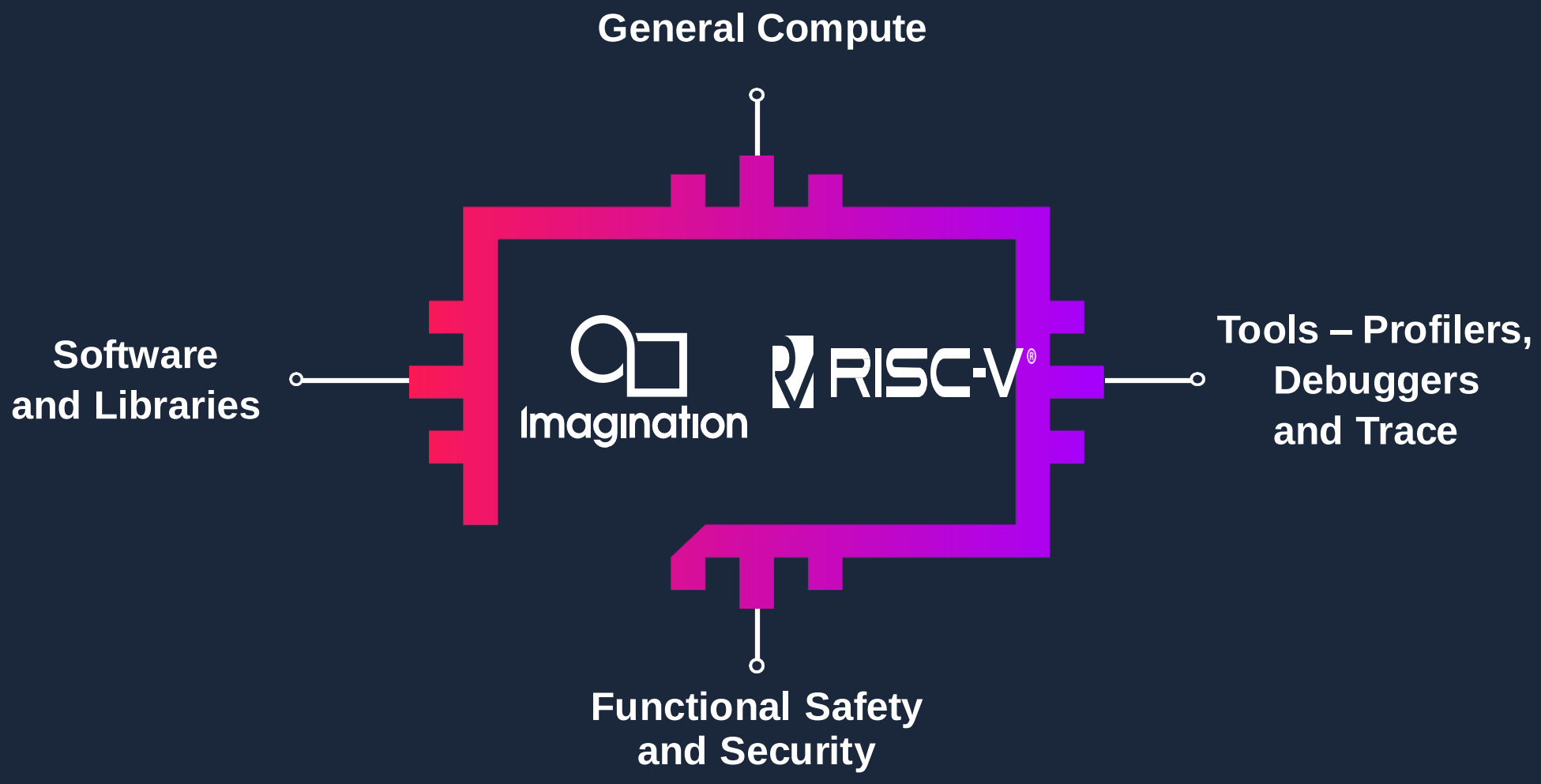
+ 低精度张量加速
+ 图形编译和显式内存管理
+ 用于图形和计算/人工智能的双用途 SRAM
+ 通过降低电压和时钟速度提高能效

消耗一些面积，获得最佳性能、功耗和面积比（PPA），并扩展加速能力。

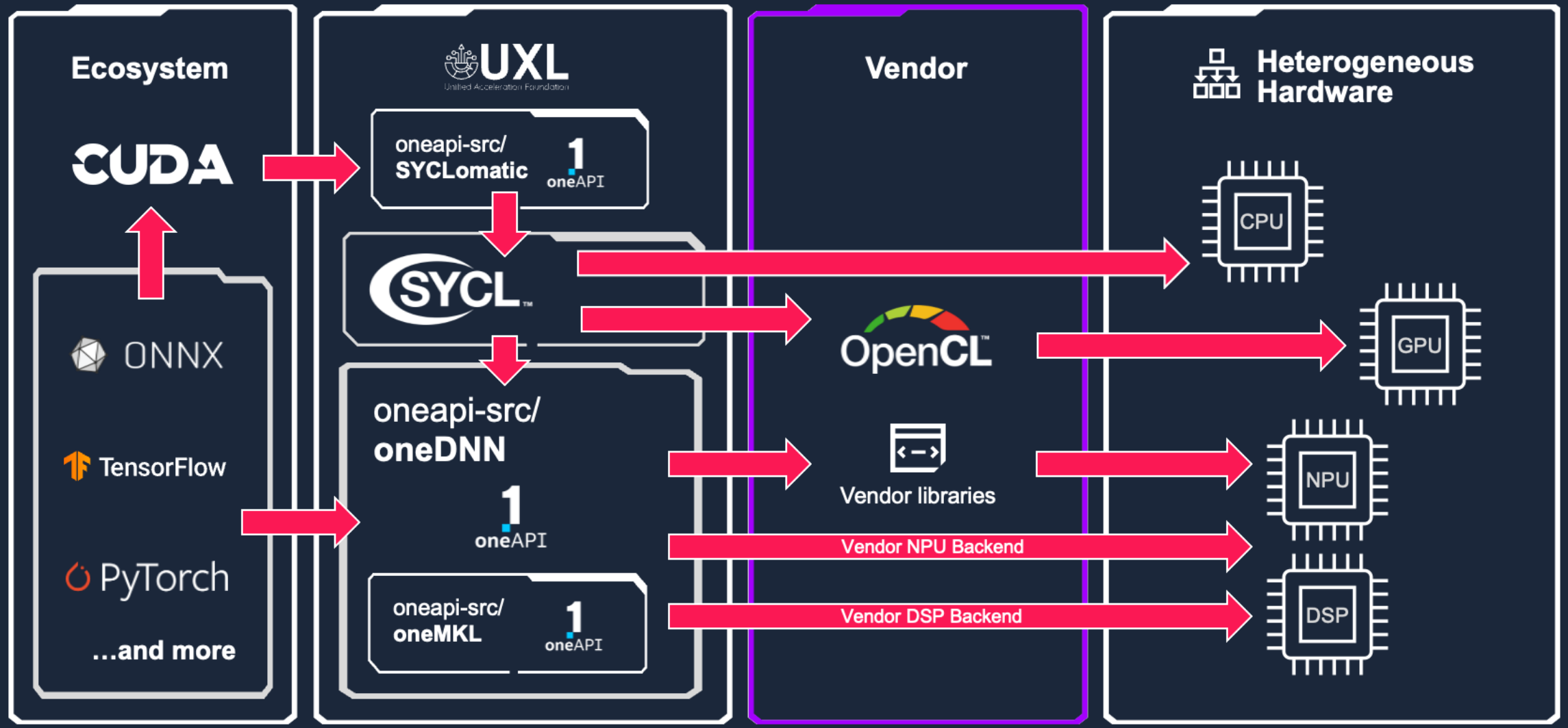
+ 更多的灵活性
+ 更广泛的张量操作菜单
+ 支持向量相似性操作
+ 更好地支持即时执行和动态张量形状
+ 更快的编译器和映射器
+ 为70亿以上参数模型重新架构
+ 支持设备上的微调

放弃一定的PPA, 获得更好的灵活性

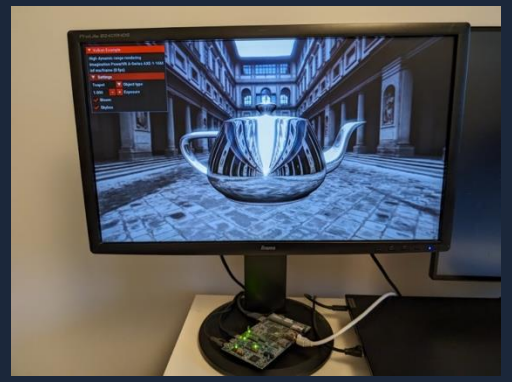
面向边缘的通用计算方案

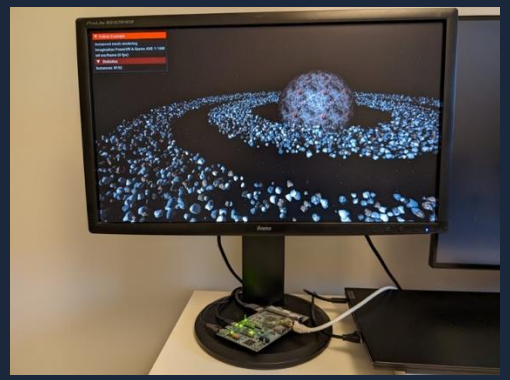


Yes, you can run it



Yes, you ALREADY can run it





OpenGL ES Vulkan EGL OpenGL OpenCL

Open Source Demo - Sascha Willems

Imagination

THANK YOU!

Please contact ke.xu@imgtec.com for any additional follow up.