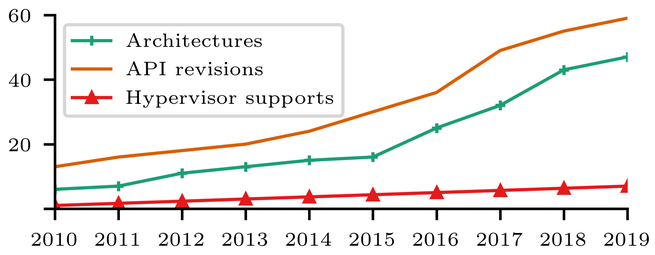
Automatic Accelerator Virtualization via Language and Compiler Support

Hangchen Yu, Arthur M. Peters, Amogh Akshintala, Christopher J. Rossbach

The University of Texas at Austin · University of North Carolina at Chapel Hill · VMware Research Group

Technology Trend



Hypervisor support falls behind accelerators' proliferation

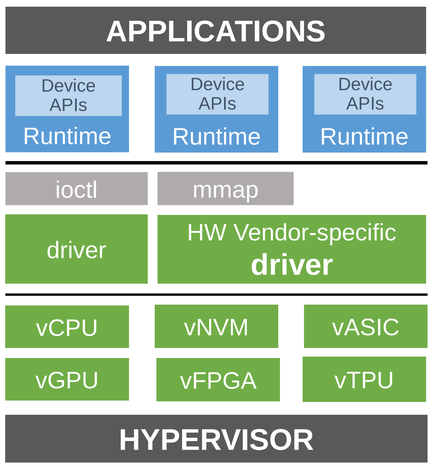
Silos Complicate Virtualization

API remoting: interposition, compatibility

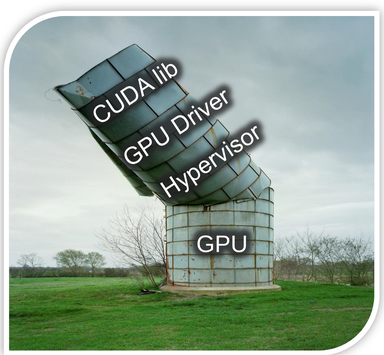
Para-virtual I/O: complexity, compatibility e.g. SVGA translates guest interactions into DirectX

Full-virtualization: significant overheads by trap-based interposition

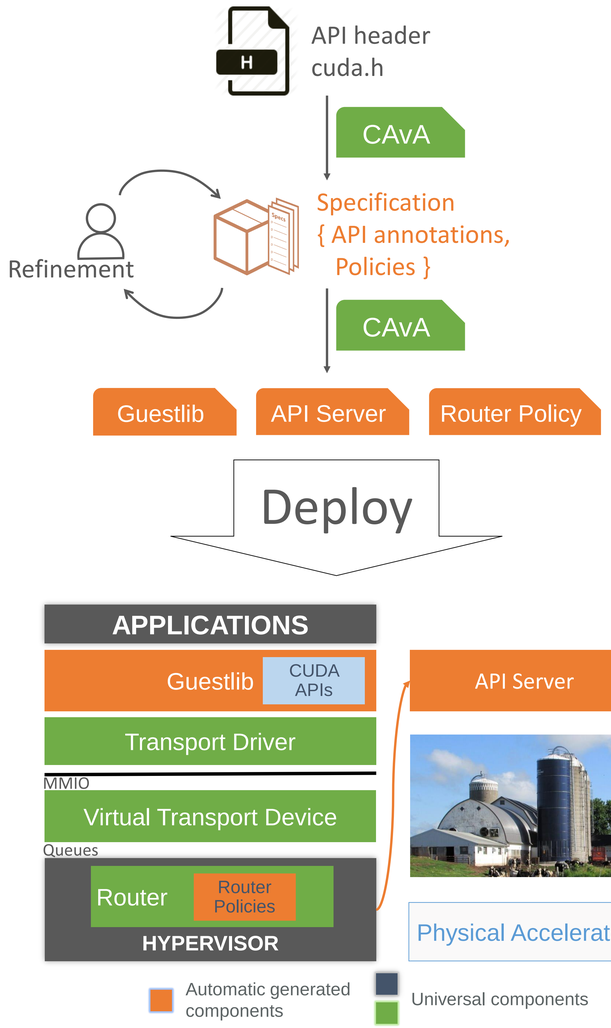
SRIOV: remains lacking by hardware support (< 0.95% NVIDIA GPUs)



Interposition **only** possible at the top or bottom of silo



AvA developer workflow



Specification Example

```
ava_throughput_resource memcpy_thp;
CUresult CUDAAPI
cuMemcpyHtoD(CUdeviceptr dstDevice,
             const void *srcHost,
             size_t byteCount) {
    ava_argument(dstDevice) ava_handle;
    ava_argument(srcHost) {
        ava_in; ava_buffer(byteCount);
    }
    ava_consumes_resource(memcpy_thp, byteCount);
}
```

Virtualized Accelerators

API	Gen	#	LoC	Hardware
OpenCL	×	39	7514	NVIDIA GTX 1080 AMD RX 580
	√	38	1060	
CUDA 10	√	16	266	NVIDIA GTX 1080
CUDART 10	√	93	1358	
TensorFlow 1.14	√	111	1865	
NCSDK v2	√	26	479	Movidius NCS
QuickAssist	√	19	444	Intel QAT 8970

……

AvA has supported **10** accelerators and **12** framework APIs

Selected Evaluation Results