

混合云算力解决方案

半导体芯片研发 如何实现降本增效

篆芯半导体案例

面临的挑战

篆芯半导体总部位于中国上海，是一家数字芯片创新企业，产品以大规模以太网交换芯片为主。在前端设计研发阶段，功能设计和验证阶段会有大量的回归验证（regression）工作，每轮回归验证（regression）会产生数千以上的并发短时验证计算任务。

篆芯在数据中心自建 HPC 集群，为芯片研发工程师提供 EDA 运行环境和算力支撑。但随着项目推进和研发团队规模的扩大，本地集群已无法满足此类突发算力需求，导致大量任务排队，影响研发效率。

但无论是购买更多物理服务器扩充线下集群，还是购买公有云资源搭建新集群，都会存在算力空闲期造成资源浪费。如何合理规划算力资源、精细平衡效率和成本，成为了产品能否按时交付的关键因素之一。

项目初期

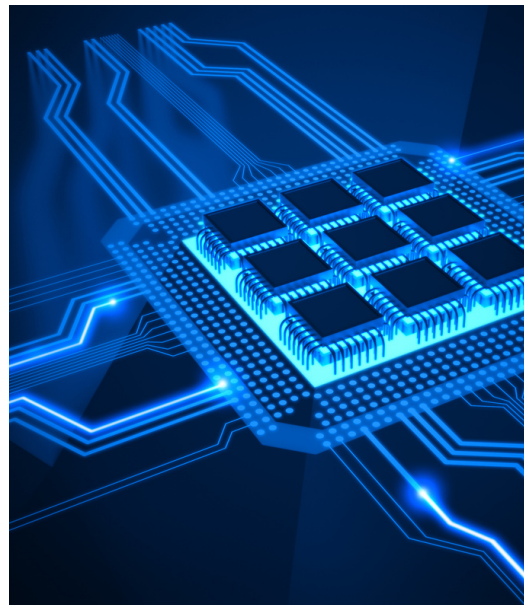
代码编写与版图设计，算力需求低

前端仿真

逻辑仿真，算力需求上升，需频繁迭代验证

后端仿真

物理设计与布局布线，算力和内存需求高峰，计算时间长



解决方案

Memory Machine Cloud (简称MMCloud) 提供本地和公有云资源的统一管理, 为使用 Synopsys、Cadence、Mentor 等主流 EDA 软件的芯片研发设计企业提供大规模弹性算力的供给, 分发, 调度, 智能匹配, 回收等全生命周期管理服务。

当线下集群资源不足时, 将计算任务自动溢出到公有云上。公有云资源完全按需申请、并按计算需求实时弹性伸缩, 最大限度减少任务排队时间。任务结束后, 计算资源自动回收, 业务闲时资源 0 浪费。



- 01 混合调度**
提供本地与云端统一智能调度、支持多维度调度策略
- 02 使用习惯**
提供 SGE/LSF/PBS/SLURM 指令, 不改变用户任务提交的使用习惯
- 03 智能算法**
基于 AI 算法智能分析任务、应用资源需求、智能匹配不同的计算资源
- 04 业务连续性**
通过断点续算能力满足任务在更换资源时, 保障任务不受影响
- 05 大规模算力**
满足大规模突发算力、智能溢出到云; 提供算力弹性伸缩、按需使用

核心技术

智能调度

自动监控云上云下资源使用情况，将任务发送到最合适的算力集群

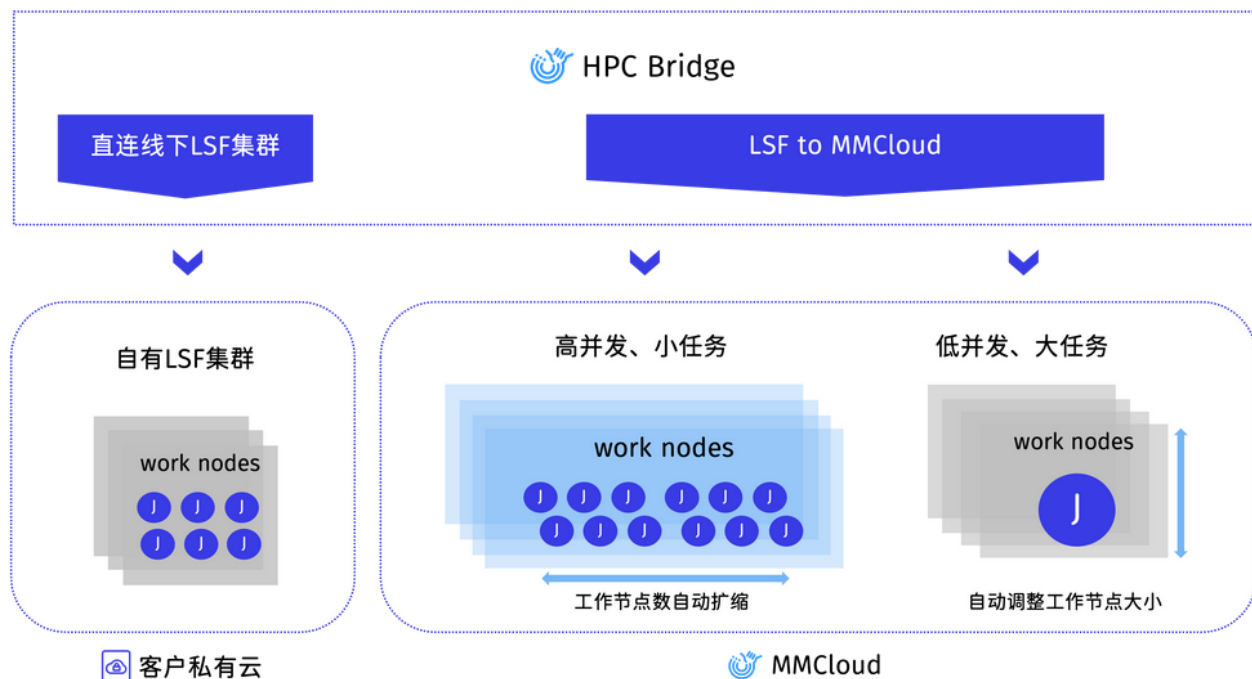
精细化调度

基于任务特性、识别任务所需资源（大任务、小任务）进行混合调度，满足高并发、高算力等不同场景需求

动态选型

对于有波峰波谷的长时任务，通过 WaveRider 特性实现计算节点的动态升降配

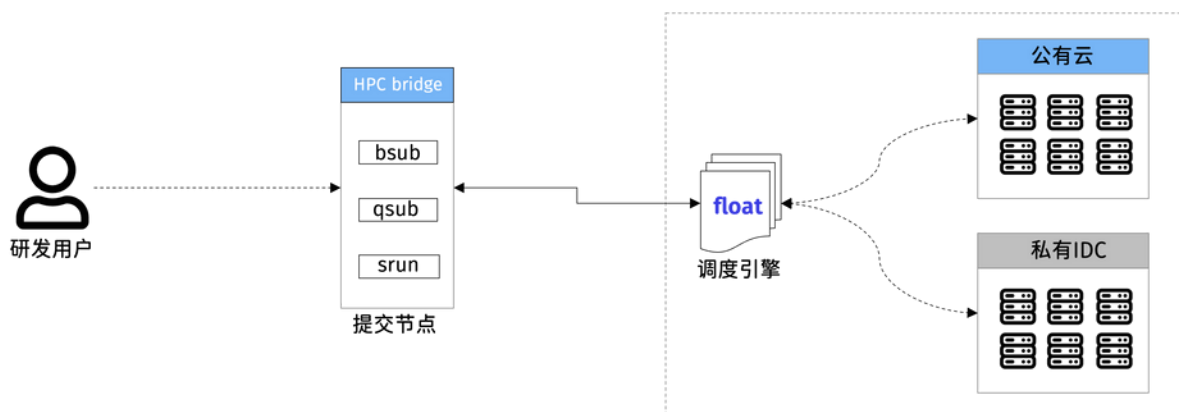
FLOAT混合调度



案例：大规模Regression降本增效

场景&需求

- 本地有小规模的算力资源，随着业务的快速发展本地资源无法满足芯片研发需求，本地资源不足导致任务排队、影响业务研发效率。
- 前端验证 regression，由于任务量大、计算时间短，业务存在波峰波谷，需要大规模并发算力提升 regression 仿真效率。
- 混合架构，本地与云端打通；本地资源不足任务自动溢出到云；计算在云端，数据在本地，提升数据安全，保障研发使用体验。



方案&效果

- 构建 IC 研发设计云平台
- 分钟级构建百万级规模算力资源
- 根据业务需求按需智能分配资源，合理使用 Spot 机型
- 根据业务运行资源消耗自动调配实例（WaveRider）
- 使用 AppCapsule 实现业务降本增效

30%
成本减少

44%
速度提升

硅谷 | 上海 | 北京 | 成都

www.mmcloud.io

sales@memverge.com

