

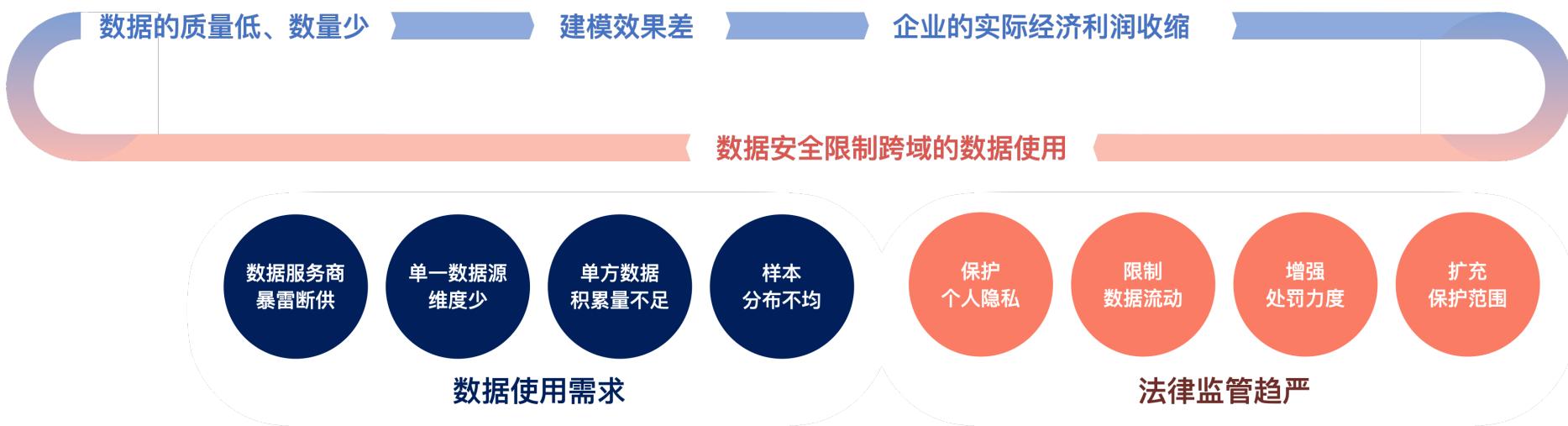
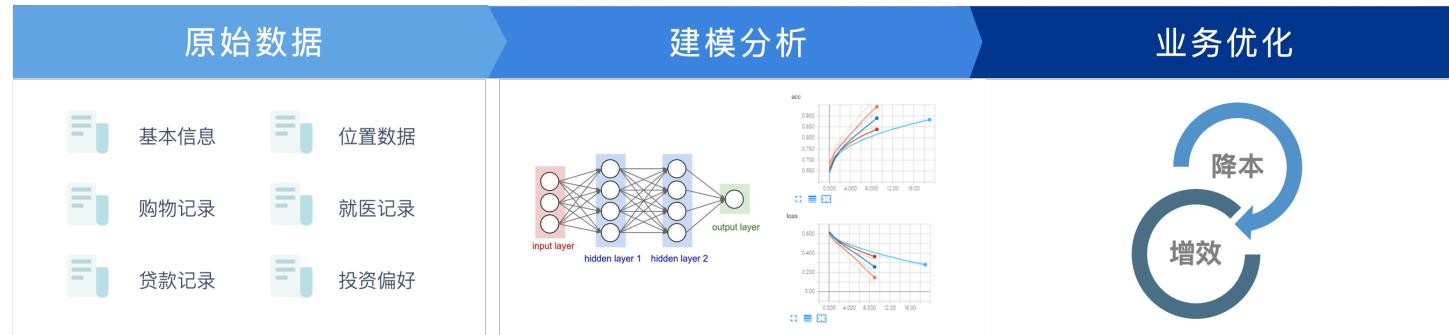


联邦学习企业级实现

张骏雪
星云Clustar CTO

张海宁
VMware中国研发技术总监

DT时代数据创造巨大价值，但隐私及安全问题日益突出



各类数据安全整个法规给行业带来机遇与挑战

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 × 云原生 × 边缘计算

《中华人民共和国个人信息保护法（草案）》 2020.10.21

第十三届全国人大常委会第二十二次会议对《中华人民共和国个人信息保护法（草案）》进行了审议，并向社会公众征求意见。

第三条 组织、个人在中华人民共和国境内处理自然人个人信息的活动，适用本法。在中华人民共和国境外处理中华人民共和国境内自然人个人信息的活动，有下列情形之一的，也适用本法：

(一)以向境内自然人提供产品或者服务为目的；(二)为分析、评估境内自然人的行为；(三)法律、行政法规规定的其他情形。

第四条 个人信息是以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，不包括匿名化处理后的信息。



个信法草案将现行的很多推荐性做法升格为强制性法定义务，所有涉及个人信息处理活动的企业都应根据规定，合法合规处理个人信息，并采取相应的管理和技术措施保护个人信息。

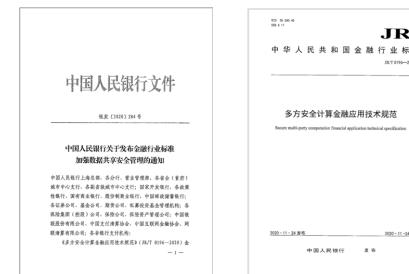
此外，个信法草案特意明文将“匿名化处理后的信息”（指个人信息经过处理无法识别特定自然人且不能复原的过程）排除在个人信息之外，将个人信息的外延进行适当限缩，给特定的信息处理活动预留一定的空间。

人工智能 × 云原生分论坛

中国人民银行关于发布金融行业标准 加强数据共享安全管理的通知 2020.11.24

《多方安全计算金融应用技术规范》（JR/T 0196-2020）金融行业标准已经全国金融标准化技术委员会审查通过，并就有关事项通知如下。

- 金融机构结合市安全集认真落实《多方安全计算金融应用技术规范》，建立数据共享机制，规范数据采集、授权、使用，确保数据专事专用、最小够用，杜绝数据被误用、滥用、博阿虎数据主体隐私不受侵害。
- 有关行业协会根据工作需要按照《多方安全计算金融应用规范》加强数据安全共享行业自律管理，建立健全自律检查、风险联防联控等机制。



《技术规范》中将MPC典型应用规范性分类为联合查询、联合建模、联合预测。举例MPC典型应用场景如基于MPC的生物特征识别、基于MPC的联合风控。

联邦学习 – 打破数据孤岛、解决人工智能落地的最后一公里

联邦学习

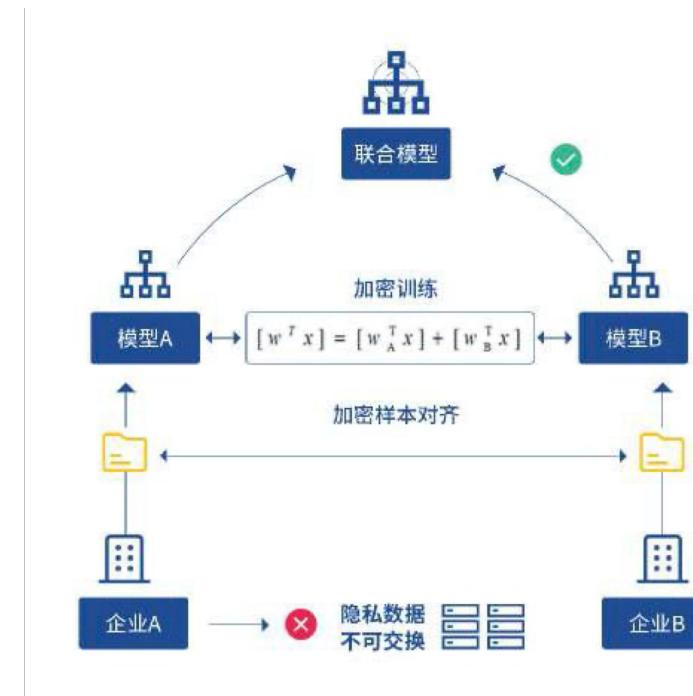
在进行机器学习的过程中，各参与方可借助其他方数据进行联合建模。各方无需共享数据资源，即在**数据不出本地**的前提下，进行**数据联合训练**，建立共享的机器学习模型。

公共价值：

- 加速人工智能技术创新发展
- 保障隐私信息及数据安全
- 促进全社会智能化水平提升

商业价值：

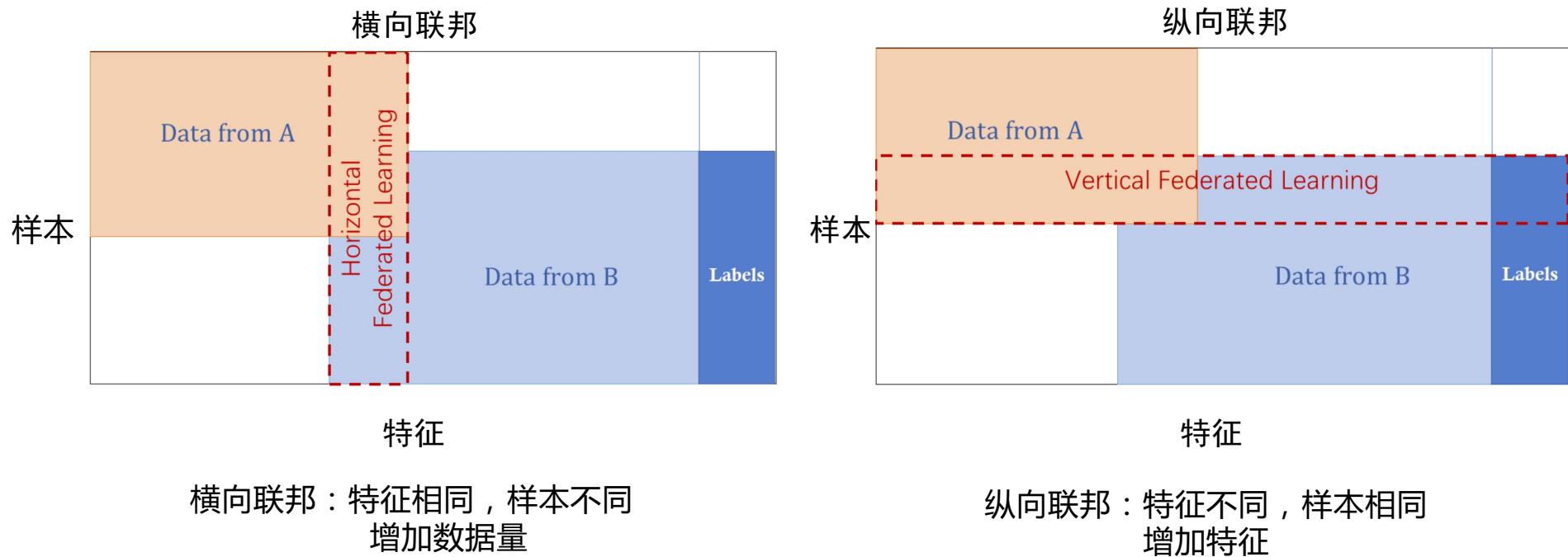
- 带动跨领域的企业级数据合作
- 催生基于联合建模的新业态和模式
- 降低技术提升成本和促进创新技术发展



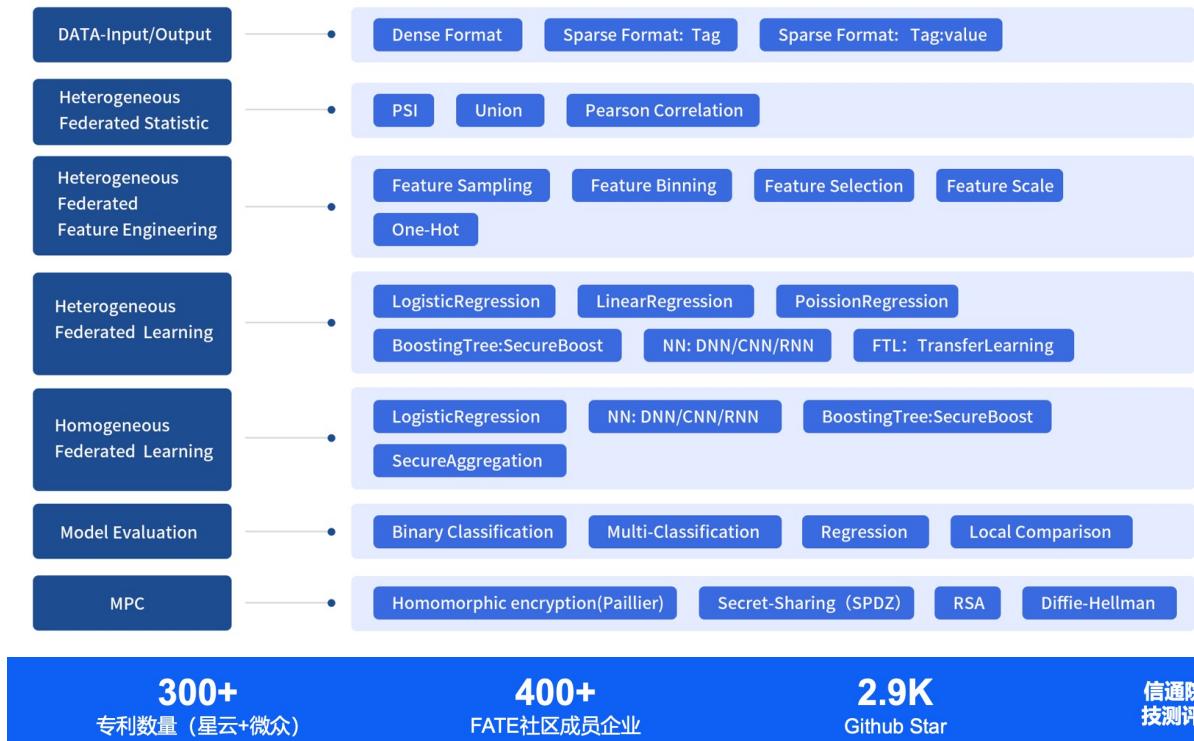
数据留在本地，参数经过**同态加密**后，在密态下进
行同步与更新，实现联合建模

联邦学习 – 打破数据孤岛、解决人工智能落地的最后一公里

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 × 云原生 × 边缘计算



FATE：国内首个开源、支持商业落地的联邦学习框架

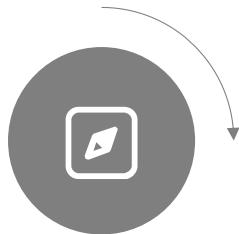


星云是FATE开源项目TSC成员之一

目前TSC成员有：

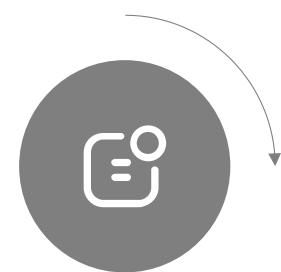
微众、VMWare、腾讯、银联、建信金科、星云、工行、中国银行、富数、中国电信等

联邦学习的几大核心问题



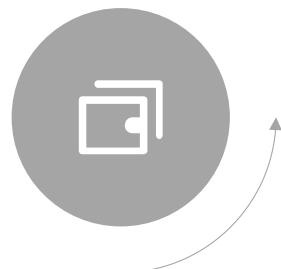
通信效率低

联邦学习往往发生在不同公司之间、需要跨数据中心进行，跨数据中心的网络带宽、延迟、稳定性都有很大局限性，造成通信效率低，通信成为联邦学习的瓶颈之一



系统易受攻击

联邦学习的最终目标是在隐私保护的前提下完成机器学习建模，数据安全与隐私是至关重要的，但在现实应用中，联邦学习系统因为是一个多方参与的分布式学习系统，非常容易遭到攻击



计算复杂度高

在采用同态加密等隐私保护技术后，隐私得到较好保护，但由于密文上计算复杂度过高，导致联邦学习效率低、耗时长。

星云在FATE中的工作 – 算力加速

GPU加速同态加密

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 × 云原生 × 边缘计算

优化策略

- 优化1: 分治做元素级并行
- 优化2: 平方乘算法+蒙哥马利算法
- 优化3: 中国剩余定理

异构加速联邦学习评测结果 – GPU vs CPU

	优化1	优化2	优化3	
同态加密/kops	 1.35x	13	 5.80x	56
同态解密/kops	 1.51x	52	 2.17x	74
密态乘法/kops	 1.22x	175	 31.4x	4522
密态加法/kops	 419x	47755		

2019 英伟达GTC大会

星云受邀在英伟达(NVIDIA)的GTC大会上发表《GPU在联邦学习中的探索》,阐述了联邦学习密态计算和密文传输的问题, 并就如何提高密态计算和密文传输的效率进行了相应的解析,深入分析联邦学习场景下, GPU计算优势与挑战, 并提出星云的优化技术。 <https://www.infoq.cn/article/VmS3QHkOIDNa3ks7yNqJ>

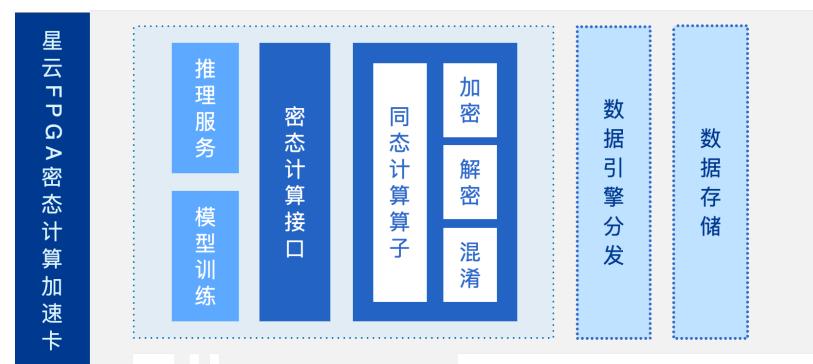
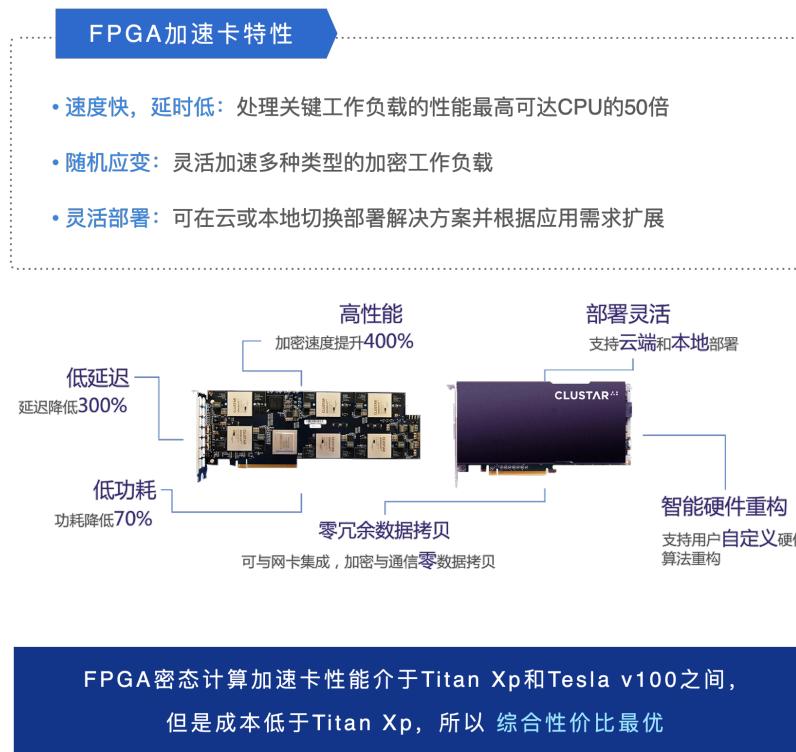


星云在FATE中的工作 – 算力加速

FPGA加速同态加密

Open Source AceCon
2021 智能云边开源峰会
人工智能 × 云原生 × 边缘计算

FPGA密态计算加速卡，是针对联邦学习等隐私计算场景，量身定制的高性能计算方案，可有效解决因使用同态加密而产生的计算压力与延时问题。



性能价格比较						
设备类型	设备名	加密性能 (kop/s)	性能比例 (/CPU)	价格 (¥ 10,000)	性能/价格	性价比
CPU	Inte Gold 6132 @ 2.60GHz, 14 core	9.23	1.00	1.50	6.15	1.00
GPU	Titan Xp	20.81	2.25	2.00	10.41	1.69
GPU	Tesla V100	52.50	5.69	5.50	9.55	1.55
FPGA	Xilinx VU13P	38.60	4.18	1.80	21.44	3.48

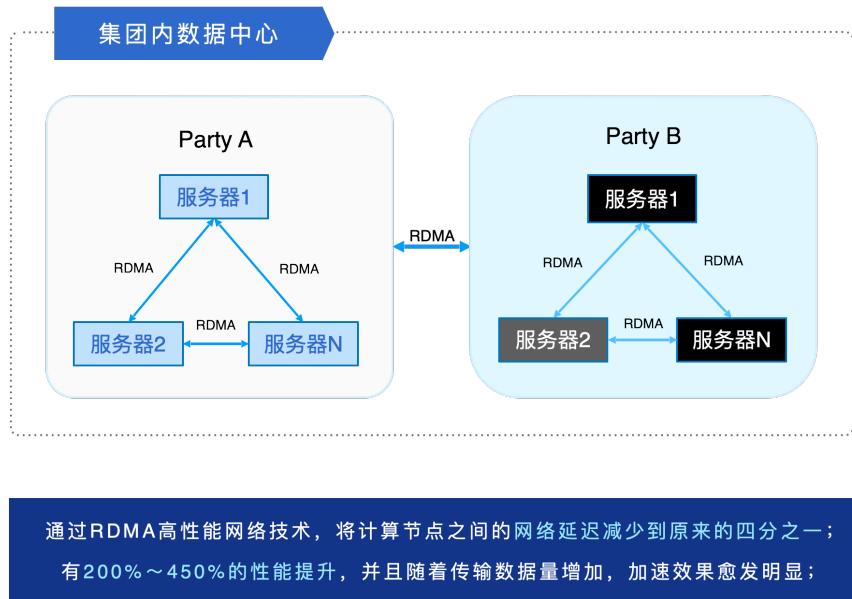
星云在FATE中的工作 – 通信加速

同数据中心联邦场景网络加速

典型场景：

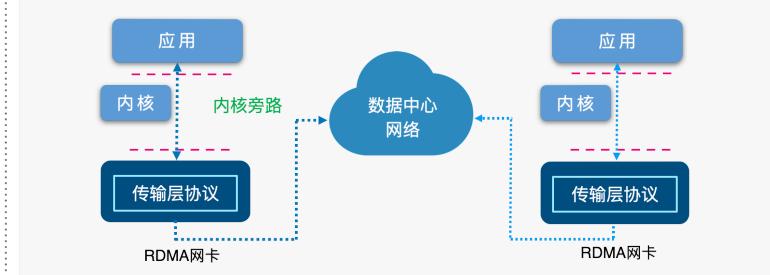
Party-A和Party-B同属一个数据中心：集团内的子公司之间联邦建模

- 单个Rollsite的多台机器之间通过RDMA进行高速通信
- 多个Rollsite的机器之间通过RDMA进行高速通信



RDMA优势

- 内核旁路(Kernel Bypass), 将传输层卸载到网卡硬件
- 高速网络下实现高吞吐、低时延、低CPU负载的两点间通信



TCP&RDMA性能对比

传输数据量	带宽 (MB/S)			延迟 (ms)		
	TCP	RDMA	加速比例	TCP	RDMA	降低到
1KB	674	1363.97	202.37%	15.3	4.17	27.25%
1MB	2150	6300.66	293.05%	584	158.71	27.18%
200MB	1570	4864.37	309.83%	144000	41442.88	28.78%
500MB	1570	4864.35	309.83%	400000	103678.31	25.92%
1GB	1070	4863.86	454.57%	1000000	212404.28	21.24%

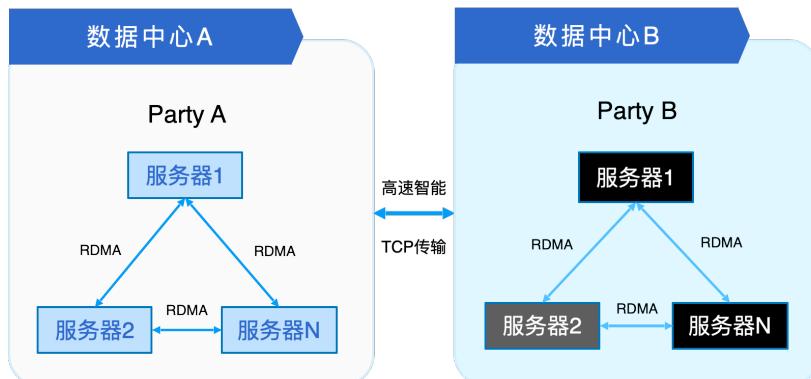
星云在FATE中的工作 – 通信加速

跨数据中心联邦场景网络加速

典型场景：

Party-A和Party-B分属不同数据中心：不同的机构之间联邦建模

- 单个Rollsite的多台机器之间通过RDMA进行高速通信
- 多个Rollsite的机器之间通过高速智能TCP传输模块进行高速通信



通过RDMA技术，解决计算节点之间的网络延迟开销，提升性能；
通过远距离通信模块优化，解决复杂网络下，网络丢包对联邦任务的影响

高速智能TCP传输

- **极高性能：**对网络丢包不敏感，比现有TCP性能有明显提升
- **高扩展性：**适配数百个linux发行版本
- **简单易用：**支持即插即用；零代价集成，应用程序无需修改；

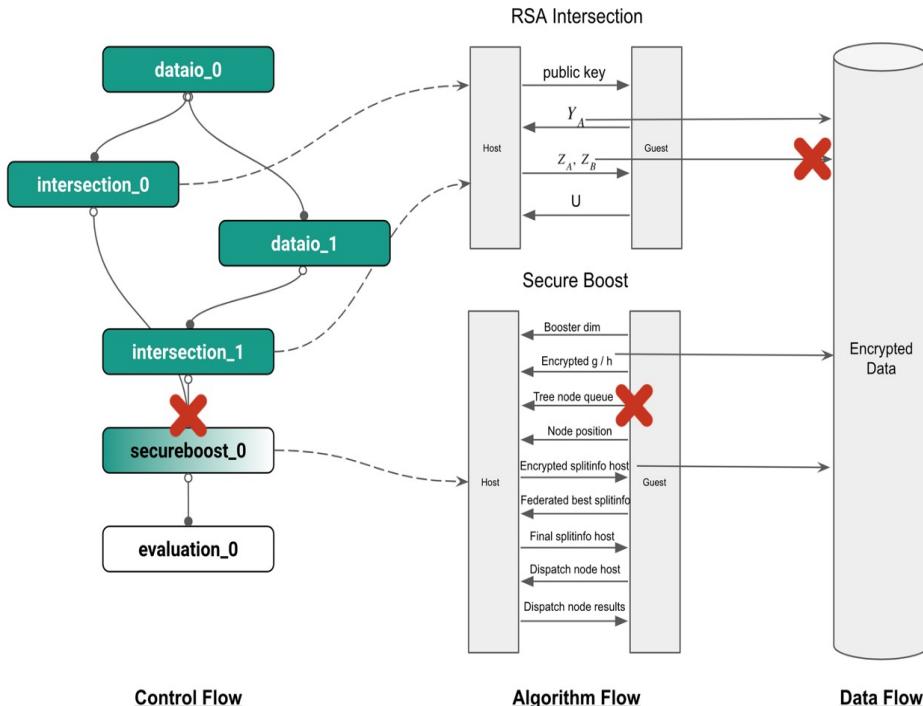
性能比较				
网络延迟 (毫秒)	丢包率 (%)	TCP (Mbps)	Pro-TCP (Mbps)	Pro-TCP/TCP
20 (相当于市内)	0.01	913.0	937	1倍
40 (相当于省内)	0.1	23.2	891	38倍
80 (相当于省际)	1	2.4	738	307倍
160 (相当于国际)	2	1.0	568	568倍
320 (相当于洲际)	4	0.5	309	618倍
320 (无线网络)	10	0.2	221	1105倍

星云在FATE中的工作 – 安全审计

基于验证的联邦学习审计系统

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 × 云原生 × 边缘计算

联邦学习面临的攻击手段



1、控制流攻击：

- a) 恶意的参与方或进程可能故意停止联邦学习任务所依赖的必需服务如数据库服务、网络连接等，使得联邦学习任务无法正常进行；
- b) 恶意的参与方或进程可能会发送恶意的控制流信息。如图所示，在运行secureboost算法时，使用其他的算法流程代替，或者替换某些算法模块。

2、算法流攻击：

- 恶意的参与方或进程可能停止联邦学习算法的执行或发送恶意的算法流信息：
- a) 破坏训练。通过停止训练，发送错误数据等方法破坏训练，导致模型失效；
 - b) 泄露隐私。通过商量好的字段等，泄露隐私信息。

3、数据流攻击

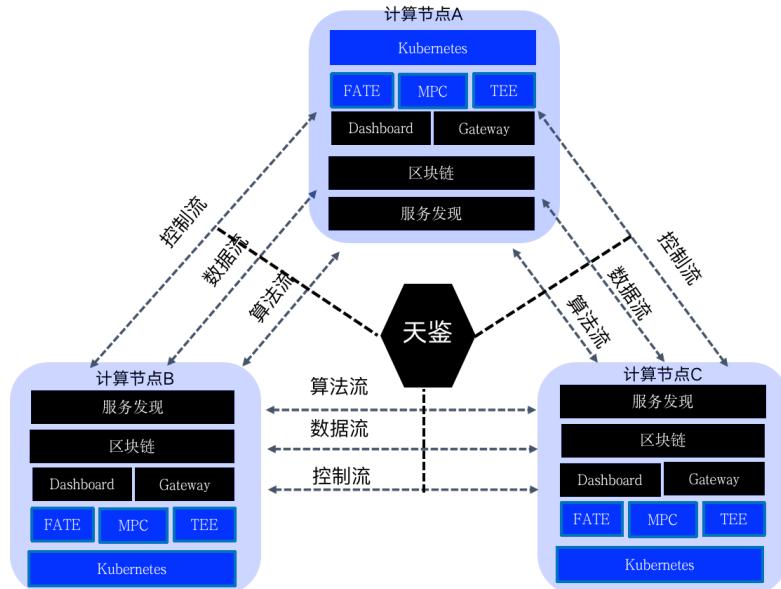
- 恶意的参与方或进程可能会发送错误或恶意的用于联邦学习的数据。
- a) 恶意或错误的数据特征，如预期的公钥位宽为1024位，而恶意进程产生的密钥位宽为128位或2025位；
 - b) 恶意或错误的数据内容：在机器学习（含联邦学习）中，每个特征都有一个合理的数据范围，恶意的数据内容可能：
 - i. 超出合理的数据范围；
 - ii. 某个必须为非空的特征为空。

星云在FATE中的工作 – 安全审计

基于验证的联邦学习审计系统

天鉴系统作为可定制的联邦学习监控报警系统，对于训练建模中传输的加密数据进行实时审计，确保源数据不会造成泄漏，在训练结束后可以再次进行事后审计，确保所有的加密数据训练都有迹可循，若造成泄漏方便追责。

- ✓ **控制流**：主要防范恶意的参与方或进程可能故意停止联邦学习任务所依赖的必需服务，如数据库服务、网络连接等。
- ✓ **算法流**：主要防范恶意的参与方或进程可能停止联邦学习算法的执行或发送恶意的算法流信息，如破坏训练、泄露隐私等。
- ✓ **数据流**：主要防范恶意的参与方或进程可能会发送错误或恶意的用于联邦学习的数据，如恶意或错误的数据特征、恶意或错误的数据内容(例如超出合理范围的数据、某个必须为非空的特征为空)等。





星云联邦学习企业级产品 x VMware VCF Tanzu 联合解决方案发布

产品核心能力

联邦学习企业级产品能力

Open Source AceCon
2021 智能云边开源峰会
人工智能 × 云原生 × 边缘计算



高安全
存证审计
权限管理
加密算法
安全链路



高性能
高性能异构算力加速
高性能网络加速
高性能优化算法
高性能计算引擎
高性能调度系统



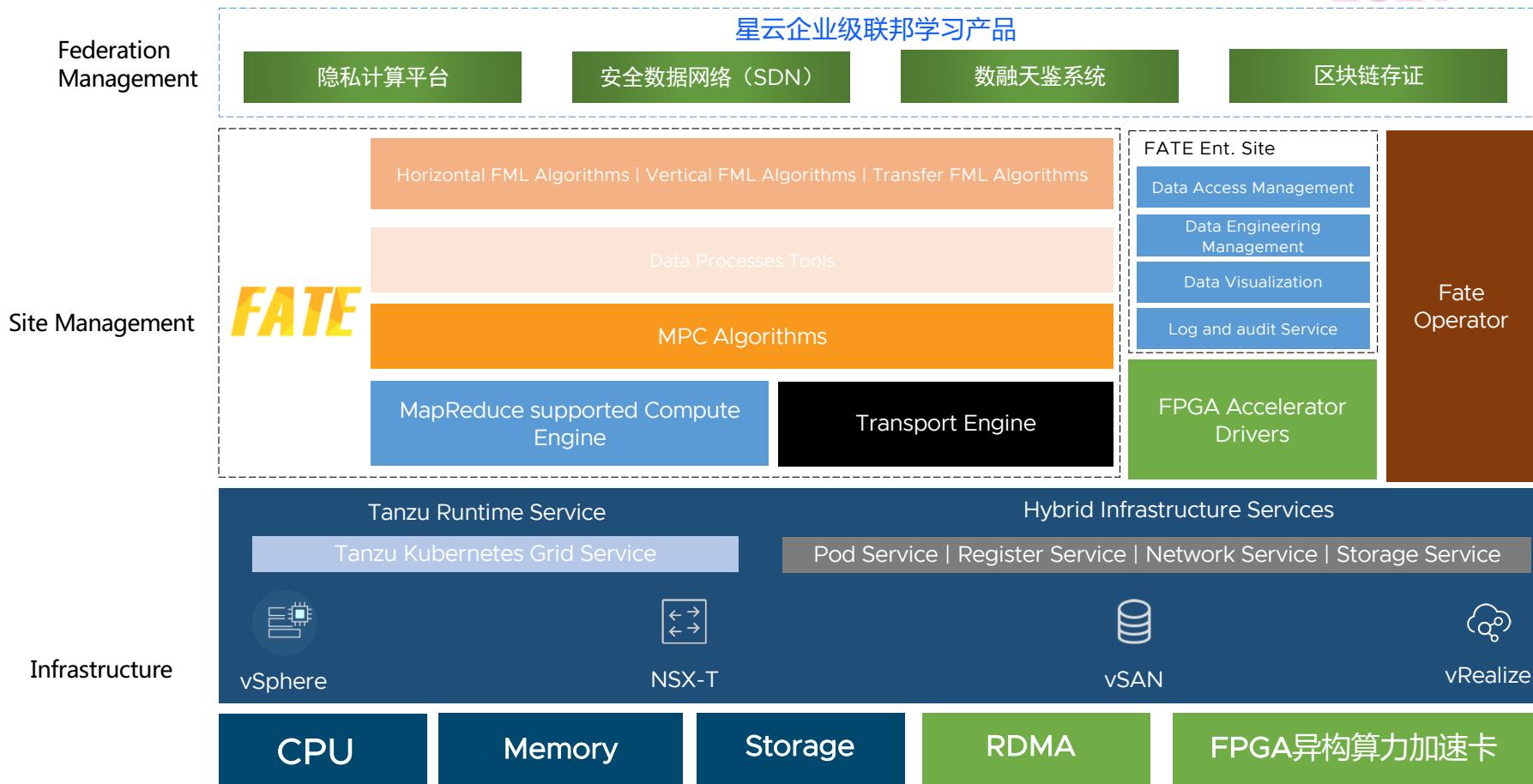
企业级
高可靠
高可用
高扩展
易用
易运维



开放性
支持开源FATE任务
支持开源机器学习平台
兼容Spark平台

星云基于 VCF Tanzu 的联邦学习企业级解决方案

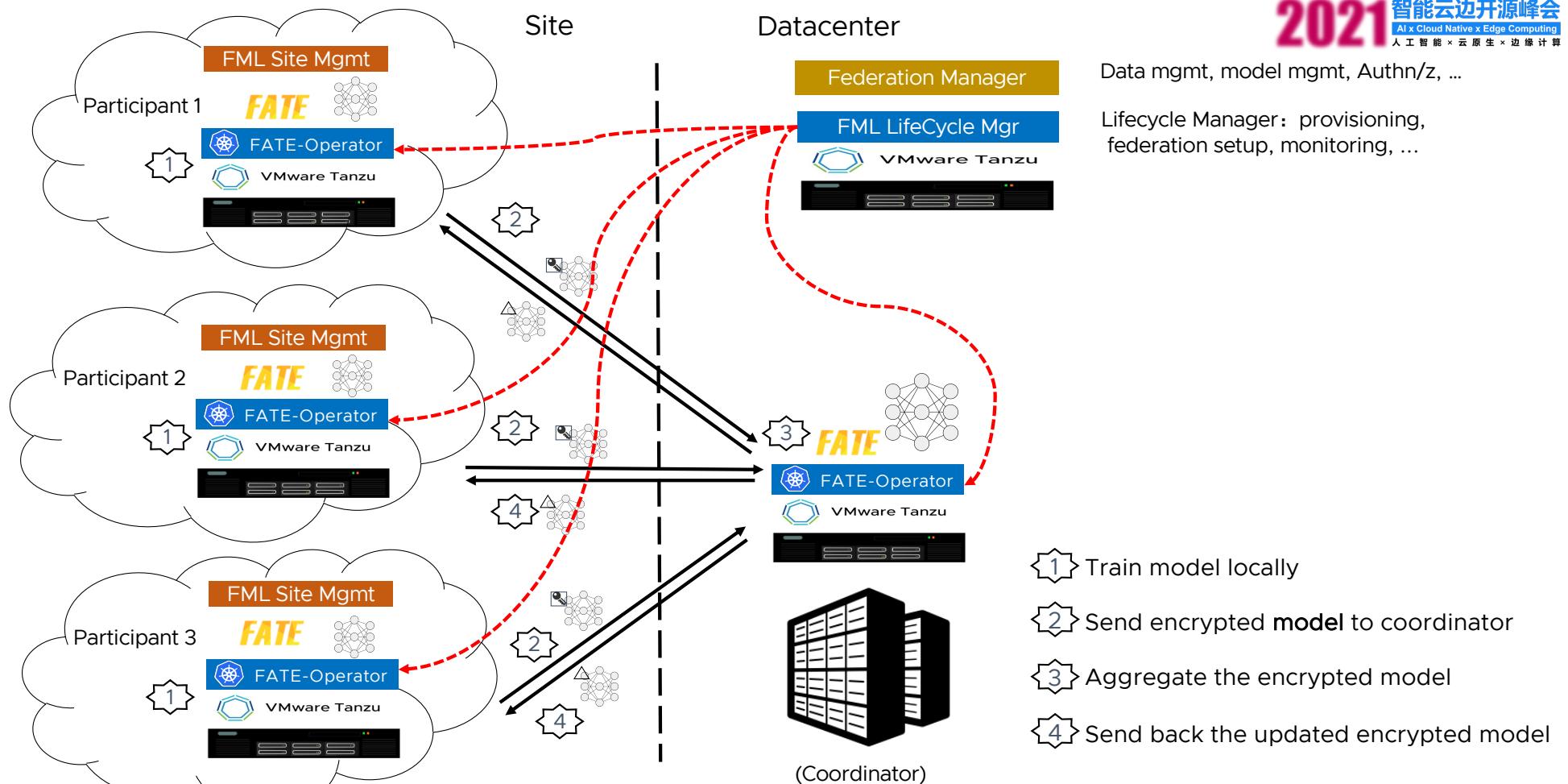
Open Source AceCon
2021 智能云边开源峰会
人工智能 × 云原生 × 边缘计算



人工智能 x 云原生分论坛

白皮书: <https://core.vmware.com/modern-kubernetes-container-applications>

联邦学习企业级解决方案管理模式



人工智能 x 云原生分论坛



Thank You