可 能 的 艺 术
THE ART OF THE POSSIBLE

第16届信息安全高级论坛

美国2024RSA参会热点研讨
INFORMATION SECURITY FORUM 2024

用主动免疫可信计算筑牢人工智能安全防线

沈昌祥

中央网信办专家咨询委员会顾问
国家集成电路产业发展咨询委员会委员
国家三网融合专家组成员

党的二十大报告提出加快建设网络强国战略任务

“没有网络安全就没有国家安全 安全是发展的前提”

第十六条 国务院和省、自治区、直辖市人民政府应当统筹规划，加大投入，扶持重点网络安全技术产业和项目，支持网络安全技术的研究开发和应用，**推广安全可信的网络产品和服务**，保护网络技术知识产权，支持企业、研究机构和高等学校等参与国家网络安全技术创新项目。

国家网络空间安全战略

夯实网络安全基础

坚持创新驱动发展，积极创造有利于技术创新的政策环境，统筹资源和力量，以企业为主体，产学研用相结合，协同攻关、以点带面、整体推进，**尽快在核心技术上取得突破**。重视软件安全，**加快安全可信产品推广应用**。

网络安全等级保护制度2.0标准及**关键信息基础设施安全保护条例**要求应当优先采购**全面使用安全可信的产品和服务**来构建关键信息基础设施安全保障体系。

PART 1

构筑安全可信人工智能新生态体系

人工智能在赋能人类社会加快发展的同时，正逐渐衍生出可危及国家安全和人类安全的重大风险：多位著名科学家警告，人工智能发展将毁灭人类社会。去年，马斯克在内的一众全球范围内AI领域重要人士表示“应将缓解人工智能导致的灭绝风险，与其它社会规模风险（如大流行病和核战争）等同重视，作为全球优先事项”

今年4月26日，美国国土安全部宣布成立一个人工智能安全与保障委员会，由国土安全部部长亚历杭德罗·马约卡斯担任主席。

ChatGPT自身无安全可信和伦理道德，一名比利时男子在与ChatGPT交流后自杀身亡。曾经马斯克为首的众位科学家向美国政府发出请愿信，要求停止ChatGPT训练，与此同时，意大利宣布禁用ChatGPT，因为 OpenAI 违反了意大利相关的隐私规则和数据保护法。

垂类大模型是大模型私有化指的是将预训练的大型人工智能模型(如GPT、BERT等)部署到企业自己的硬件环境或私有云平台上。私有化部署能够给企业带来数据安全性和自主控制能力。但是安全风险仍然存在。

人工智能的潜在风险

智能硬件被恶意用于恐怖袭击

智能化武器与军备竞赛引发国际担忧

模型失误会给人类带来灾难

智能工具被用于干预舆论和引导政治走向

机器人智能行为体一旦失控将危及人类安全

网络空间面临严重威胁

1. 2017年5月12日爆发的“WannaCry”的勒索病毒，通过将系统中数据信息加密，使数据变得不可用，借机勒索钱财。病毒席卷近150个国家和地区，教育、交通、医疗、能源网络成为本轮攻击的重灾区。

2. 2018年8月3日，台积电遭到勒索病毒入侵，几个小时之内，台积电在中国台湾地区的北、中、南三个重要生产基地全部停摆，造成约十几亿美元的营业损失。

3. 2021年5月7日，美国最大的成品油管道运营商Colonial Pipeline受到勒索软件攻击，被迫关闭其美国东部沿海各州供油网络，美国政府宣布美国17个州和华盛顿特区进入紧急状态。

认清网络安全实质

网络安全实质 ： 风险度 = 脆弱度 [ ] 威胁度

脆弱度↓

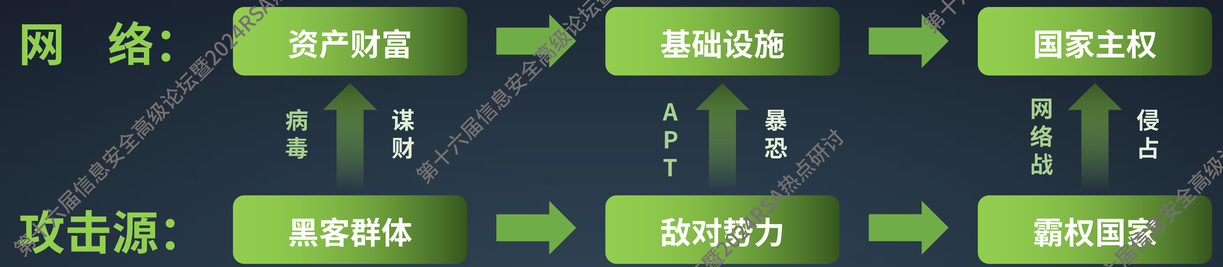
网络空间极其脆弱

是

- 计算科学问题 ---------- 图灵计算原理（少攻防理念）
- 体系结构问题 ---------- 冯诺伊曼架构（缺防护部件）
- 计算模式问题 ---------- 重大工程应用（无安全服务）

可 能 的 艺 术
THE ART OF THE POSSIBLE
第16届信息安全高级论坛
美国2024RSA参会热点研讨
INFORMATION SECURITY FORUM 2024

威胁度↑



永远命题

设计IT系统不能穷尽所有逻辑，利用逻辑缺陷挖掘漏洞，进行攻击的风险始终存在，传统“封堵查杀”难以应对未知恶意攻击。

安全可信

降低脆弱性，用安全可信产品和服务，在计算同时并行进行动态的全方位整体防护，使得完成计算任务的逻辑组合不被篡改和破坏，达到预期的计算目标。相当于人体具有免疫力确保健康。

战略任务

按国家网络安全法律、战略及等级保护制度要求用安全可信网络产品和服务构建主动免疫防护保障体系。

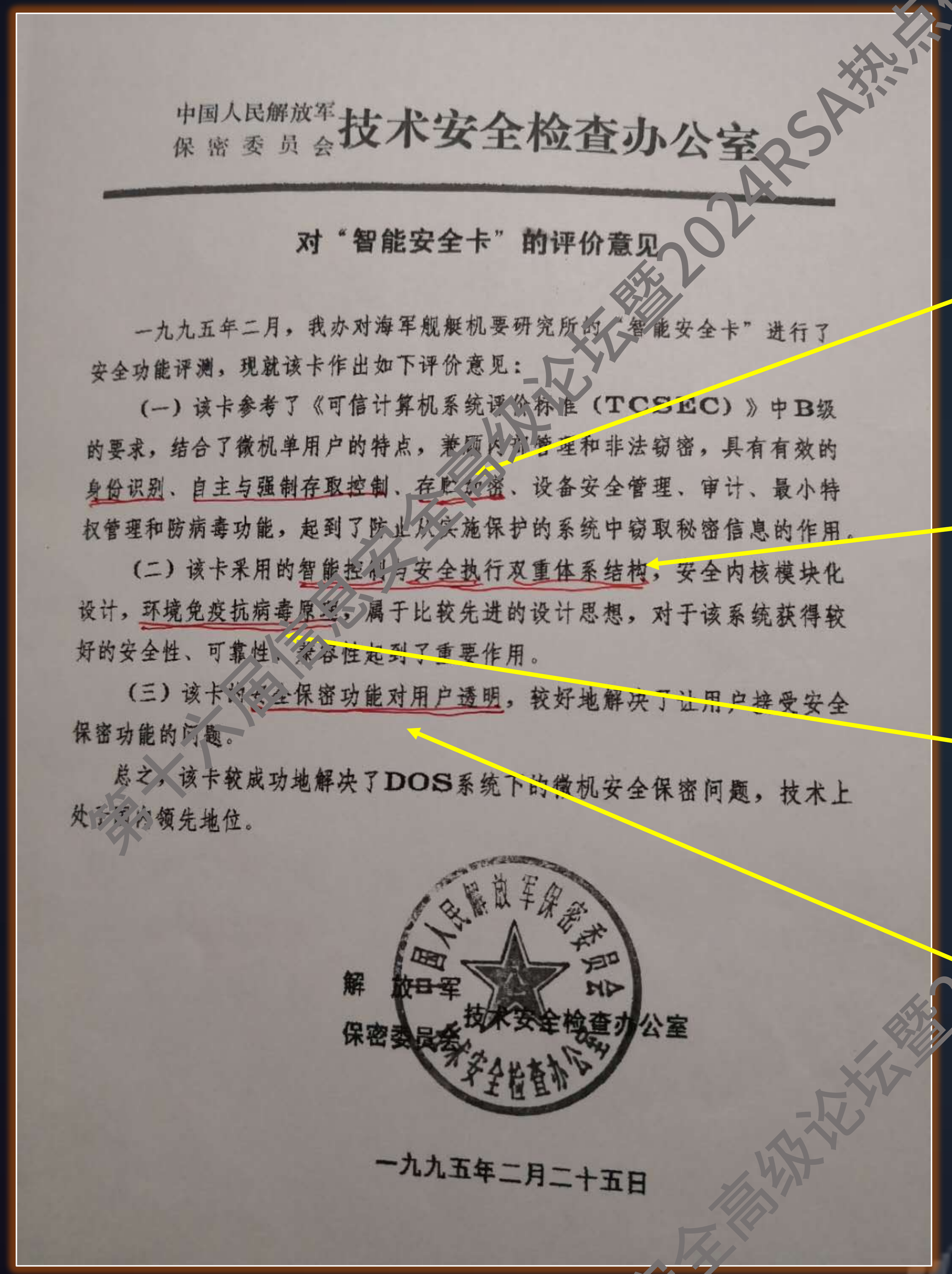
“六不”防护效果

“WannaCry”、“Mirai”、“黑暗力量”、“震网”、“火焰”、“心脏滴血”等不查杀而自灭

2 PART

构建主动免疫人工智能安全产业空间

开创可信计算3.0时代


中国人民解放军保密委员会技术安全检查办公室

对"智能安全卡"的评价意见

一九九五年二月，我办对海军舰艇机要研究所的"智能安全卡"进行了安全功能评测，现就该卡作出如下评价意见：

（一）该卡参考了《可信计算机系统评价标准（TCSEC）》中B级的要求，结合了微机单用户的特点，兼顾内部管理和非法窃密，具有有效的身份识别、自主与强制存取控制、存贮加密、设备安全管理、审计、最小特权管理和防病毒功能，起到了防止从实施保护的系统中窃取秘密信息的作用。

（二）该卡采用的智能控制与安全执行双重体系结构，安全内核模块化设计，环境免疫抗病毒原理，属于比较先进的设计思想，对于该系统获得较好的安全性、可靠性、兼容性起到了重要作用。

（三）该卡的安全保密功能对用户透明，较好地解决了让用户接受安全保密功能的问题。

总之，该卡较成功地解决了DOS系统下的微机安全保密问题，技术上处于国内领先地位。

解放军
保密委员会技术安全检查办公室

一九九五年二月二十五日

- 公钥密码身份识别、对称密码加密存储
- 智能控制与安全执行双重体系结构
- 环境免疫抗病毒原理
- 数字定义可信策略对用户透明

中国可信计算源于1992年立项研制免疫的综合安全防护系统（智能安全卡），于1995年2月底通过测评和鉴定。经过长期军民融合攻关应用，形成了自主创新安全可信体系，开启了可信计算3.0时代。

用可信计算构筑网络安全

■ 中国工程院院士 沈昌祥

当前，网络空间已经成为继陆、海、空、天之后的第五大主权领域空间，是国际战略竞争的新领域的演进，对我国网络安全提出了严峻的挑战。习近平总书记强调，建设网络强国，要有自己的技术，有过硬的技术。解决信息化核心技术设备受制于人的问题，需要从计算模式和体系结构上创新驱动。创新发展可信计算技术，推动其产业化，是将我国建设成为"技术先进、设备领先、攻防兼备"网络强国的重要举措。

一、可信可用方能安全交互

网络空间的安全与人类社会休戚相关。在人类社会中，信任是人们相互合作和交往的基础，如果我们确定对方不可信，就不会与其合作和交往。网络空间由于其开放性，允许两个网络实体未经过任何事先的安排或资格审查，就可以进行交互。这就导致我们在进行交互时有可能对对方实体一无所知。对方实体可能是通

求是杂志 2015·20

可信可用方能安全交互　主动免疫方能有效防护　自主创新方能安全可控

沈昌祥：
可信计算让信息系统国产化真正落地

Shen Changxiang: Trusted Computing Ensure That The Information System Localization Takes Effect

新华社《中国名牌》　可信计算：网络安全的主动防御时代

世界可信计算演进

可信1.0（主机）

主机可靠性
计算机部件
冗余备份
故障诊查
容错算法

世界容错组织为代表

可信2.0（PC）

节点安全性
PC单机为主
功能模块
被动度量
TPM+TSS

TCG为代表

TPM受侧信道攻击危及全球十几亿节点

可信3.0（网络）

公钥、对称双密码主动系统免疫
终端、服务器、存储系统体系可信
宿主+可信双节点平行架构
基于网络可信服务验证
动态度量实时感知

中国为代表

容错组织 — TCSEC →TCG — 中国可信计算创新

完备的可信计算3.0产品链，将形成巨大的新型产业空间

2020年10月28日

国家等级保护2.0与可信计算3.0攻关示范基地

成立揭牌

具备可信计算功能的国产CPU

嵌入式可信芯片及可信根

具备可信计算3.0技术的设备

等保2.0新标准把云计算、移动互联网、物联网和工控等采用可信计算3.0作为核心要求，筑牢智能网络安全防线

	一级	二级	三级	四级
等级保护标准可信计算要求	所有计算节点都应基于可信根实现开机到操作系统启动的可信验证。	所有计算节点都应基于可信根实现开机到操作系统启动，再到应用程序启动的可信验证。并将验证结果形成审计记录。	所有计算节点都应基于可信根实现开机到操作系统启动，再到应用程序启动的可信验证，并在应用程序的关键执行环节对其执行环境进行可信验证，主动抵御入侵行为。并将验证结果形成审计记录，送到管理中心。	所有计算节点都应基于可信计算技术实现开机到操作系统启动，再到应用程序启动的可信验证，并在应用程序的所有执行环节对其执行环境进行可信验证，主动抵御入侵行为。并将验证结果形成审计记录，送到管理中心，进行动态关联感知，形成实时的态势。

可信宿主	TCM	TPCM	检验软件	可信软件基（TSB）		
	静态可信验证基础软件可信			建链检验 应用程序可信	动态度量 执行环境	实时感知 关联态势
	BIOS	引导OS，装载系统		应用加载	应用执行	所有执行
	一级			二级	三级	四级

安全可信人工智能

人工智能大模型安全可信：

物理环境信息数据

算法模型规则及伦理

实现软件代码和流程

计算处理服务资源安全可信：

传感控制设备

通信网络

计算能力（云）

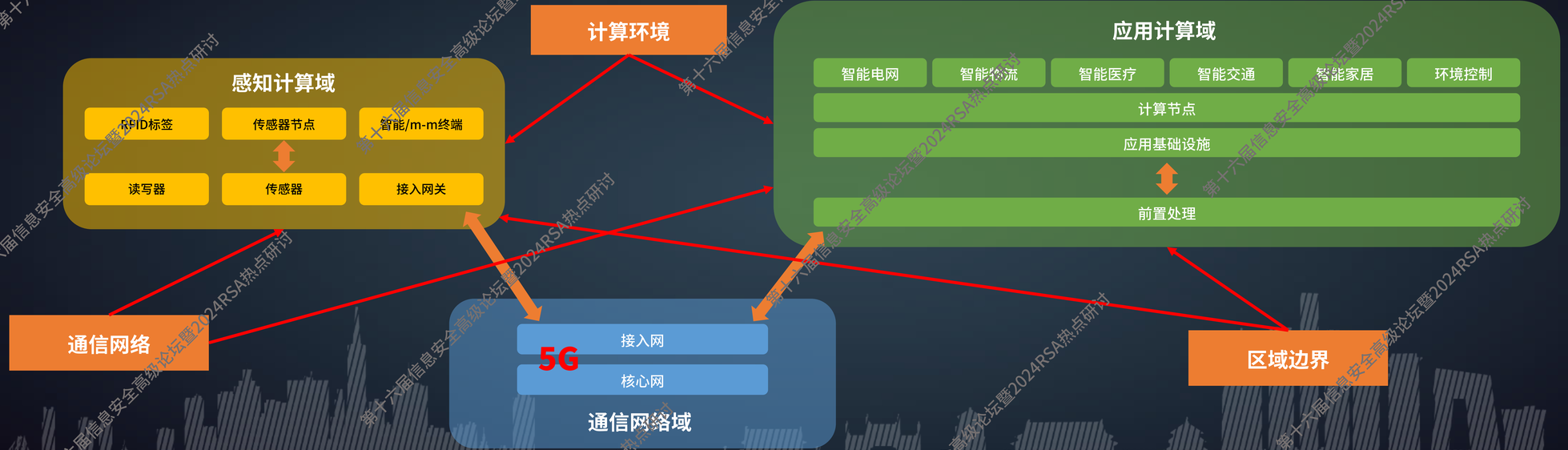
安全可信管理：

系统资源配置满足高强度计算

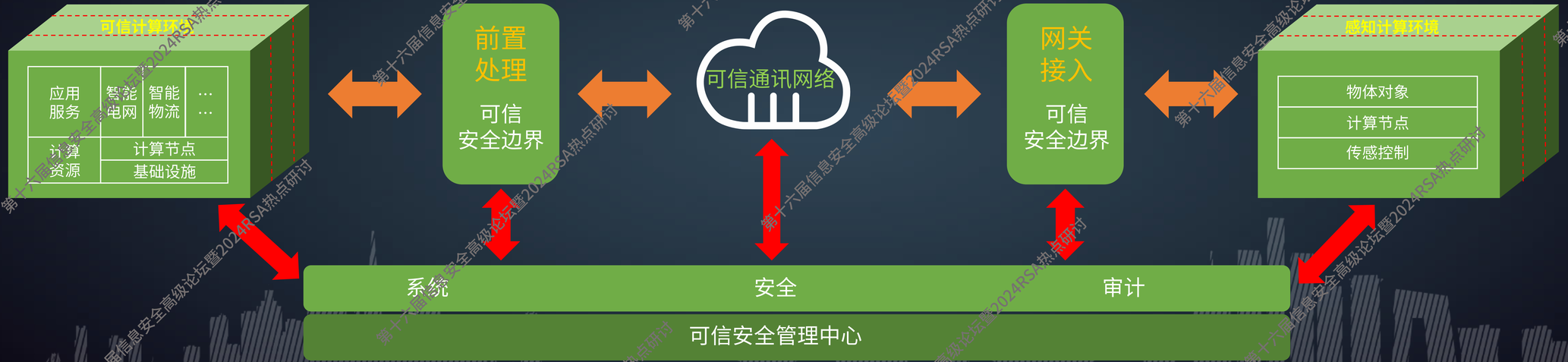
可信策略行为准则

异常控制与审计监控

智能物联网安全框架



物联网安全框架（安全管理中心支持下的三重防御）



可 能 的 艺 术

THE ART OF THE POSSIBLE

第16届信息安全高级论坛

美国2024RSA参会热点研讨

INFORMATION SECURITY FORUM 2024

感谢聆听！