

LUIS TOSSI

DA ERA DO MAINFRAME ÀS AI FACTORIES

*A evolução da TI, da computação corporativa centralizada à infraestrutura
industrial de inteligência artificial*

SÃO PAULO
2026

RESUMO

A evolução da tecnologia da informação pode ser interpretada como uma sequência de ondas de industrialização da capacidade computacional. Cada onda nasce de uma limitação da anterior, atinge escala de mercado quando encontra padronização técnica, ecossistema de fornecedores e valor econômico claro, chega à maturidade quando se torna previsível e amplamente adotada, e entra em declínio relativo quando deixa de ser a plataforma dominante para novas cargas. O declínio, neste artigo, não significa desaparecimento. Mainframes, PCs, redes locais e data centers próprios continuam relevantes. O que muda é o centro de gravidade do investimento, da inovação e do crescimento do mercado.

A sequência analisada é: mainframe, microprocessadores e PC/x86, redes locais e cliente-servidor, Internet comercial, virtualização, cloud pública, colocation moderno, local zones e edge, hyperscale, IA generativa e AI factories. A leitura executiva é que a TI deixou de ser apenas suporte administrativo para tornar-se infraestrutura produtiva.

Este artigo dedica atenção ampliada à cloud e aos data centers porque a nuvem não é um fenômeno abstrato. Ela depende de prédios, energia, refrigeração, conectividade, cadeia de suprimentos, operação 24x7 e contratos de longo prazo. A ascensão dos hyperscalers e das AI factories torna essa materialidade ainda mais explícita.

A tese central é que as AI factories representam a nova fronteira da infraestrutura digital: instalações e plataformas que transformam dados, energia, chips, modelos e engenharia em inteligência. Essa transformação reposiciona data centers como ativos comparáveis a fábricas, refinarias, portos, redes elétricas e demais infraestruturas críticas da economia contemporânea.

Palavras-chave: tecnologia da informação; data centers; computação em nuvem; inteligência artificial; AI factories.

Sumário executivo das ondas tecnológicas

Onda	Escala de mercado	Maturidade	Declínio relativo	Nova onda dominante
Mainframe	1964–1970	1970–1985	Declínio relativo a partir dos anos 1990; permanece crítico em bancos, governo e transações	PC/x86, cliente-servidor e depois cloud
PC/x86	1981–1990	1990–2005	Declínio como plataforma estratégica isolada; permanece como endpoint	Web, mobilidade, SaaS e cloud
Redes locais / cliente-servidor	1985–1995	1995–2005	Declínio do modelo departamental isolado; permanece em ambientes corporativos	Internet, web apps, SOA e cloud
Internet / Web	1993–2000	2000–2015	Não declina como base; declina a web monolítica centralizada em servidor único	Cloud, mobile, plataformas digitais
Virtualização	1999–2008	2008–2015	Declínio relativo da VM como unidade única de modernização	Containers, cloud nativa, serverless
Cloud pública	2006–2015	2015–presente	Não declina; amadurece e se especializa	Multi-cloud, hybrid cloud, local zones, IA
Colocation moderno	1998–2010	2010–presente	Não declina; reposiciona-se como interconexão e alta densidade	Cloud on-ramps, hyperscale, IA
Local zones / edge	2019–2025	Em maturação	Ainda não há declínio; mercado seletivo por latência e soberania	Arquiteturas distribuídas de IA
Hyperscale	2010–2018	2018–presente	Não declina; sofre pressão por energia, GPUs e regulação	AI factories e cloud acelerada
IA generativa	2022–2025	Em rápida maturação	Sem declínio; ciclos de consolidação de modelos e custos	Agentes, IA multimodal, AI factories
AI factories	2023–2026	Emergente	Sem declínio; fase inicial industrial	Infraestrutura energética-computacional especializada

O quadro deve ser lido como mapa de substituição parcial, não como linha de extinção. O mercado raramente abandona integralmente uma plataforma. Ele desloca crescimento, orçamento e inovação para uma nova camada, enquanto a camada anterior se estabiliza, torna-se especializada ou passa a ser integrada por APIs.

1. Introdução: a TI como infraestrutura produtiva

Durante décadas, a tecnologia da informação foi descrita como área de suporte. Essa descrição já não é suficiente. Bancos digitais, varejo omnichannel, streaming, logística, saúde, pesquisa científica, governo eletrônico, manufatura avançada e IA generativa dependem de infraestrutura digital permanente. Assim como a indústria do século XX dependia de aço, energia, transporte e máquinas, a economia do século XXI depende de chips, data centers, redes, software, modelos de IA e talento técnico.

A evolução histórica mostra uma alternância entre centralização e descentralização. O mainframe centralizou processamento e governança. O PC descentralizou produtividade. As redes locais conectaram departamentos. A Internet conectou organizações e mercados. A virtualização recentralizou servidores em ambientes mais eficientes. A cloud industrializou a infraestrutura como serviço. O hyperscale concentrou escala global em poucos provedores. A IA, finalmente, transforma computação em capacidade cognitiva. (HENNESSY; PATTERSON, 2019; REED; GANNON; DONGARRA, 2022)

A pergunta executiva central deixa de ser "qual servidor comprar?" e passa a ser "qual capacidade digital precisamos controlar, contratar ou distribuir para competir?". O artigo organiza essa narrativa com foco na infraestrutura, pois é nela que a transformação digital se materializa.

2. Metodologia e base de análise

A análise combina narrativa histórica, revisão bibliográfica e interpretação executiva. A avaliação de cada onda tecnológica usa quatro critérios: início da escala de mercado, maturidade, declínio relativo e nova onda substituta.

Essa metodologia evita dois erros comuns: tratar tecnologia como moda linear, em que uma inovação elimina completamente a anterior, e tratar infraestrutura como commodity neutra. Na prática, cada camada herdada continua influenciando requisitos, cultura operacional, segurança, custos e decisões futuras.

3. Mainframes: a computação corporativa centralizada

Início em escala de mercado	1964–1970, com a família IBM System/360 e a consolidação de ambientes corporativos transacionais.
Período de maturidade	1970–1985, quando bancos, governos, seguradoras, companhias aéreas e grandes corporações padronizaram processamento centralizado.
Declínio relativo	A partir dos anos 1990, quando servidores x86, cliente-servidor e aplicações distribuídas passaram a capturar novos projetos. O mainframe não desapareceu; tornou-se plataforma especializada.
Onda que absorveu o crescimento	PC/x86, redes cliente-servidor, servidores distribuídos e, posteriormente, cloud híbrida.



O mainframe: a arquitetura que consolidou o processamento centralizado a partir de 1964

Os mainframes surgiram como resposta à necessidade de processar grandes volumes de dados com confiabilidade. A família IBM System/360, lançada em 1964, foi decisiva porque ofereceu uma arquitetura comum em máquinas de diferentes capacidades, reduzindo o risco de migração e protegendo investimentos em software. (EVANS, 1986; PUGH; JOHNSON; PALMER, 1991) (PUGH; JOHNSON; PALMER, 1991)

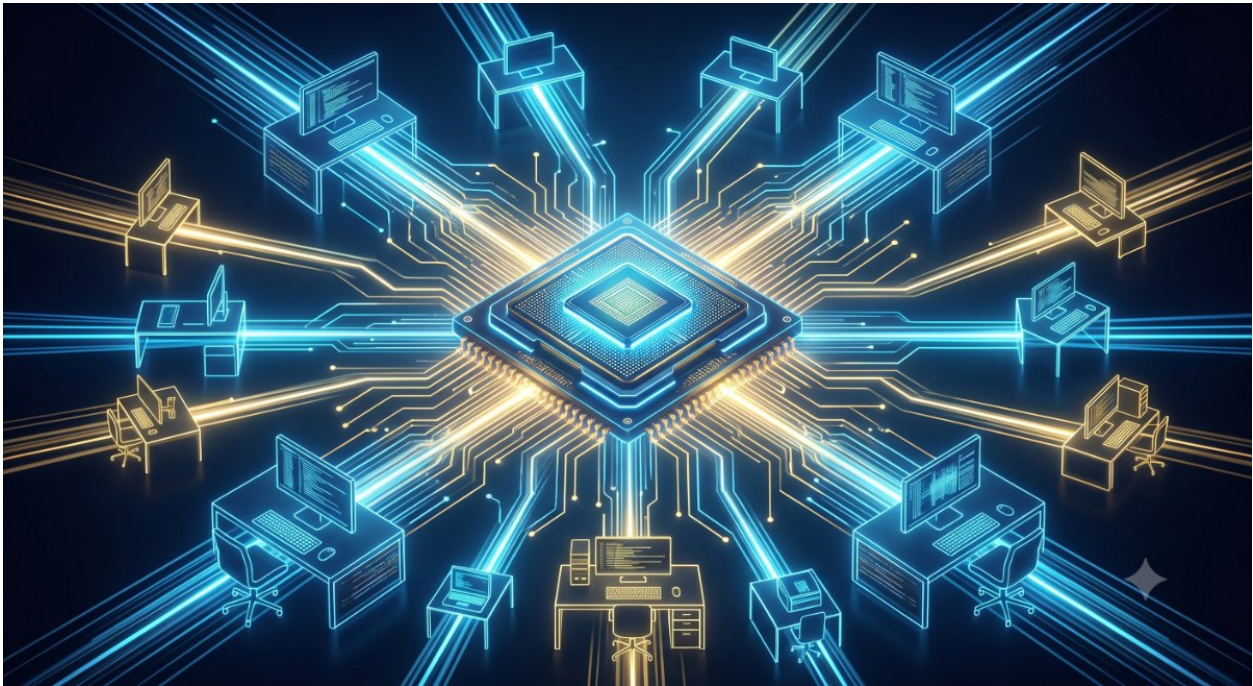
Em termos operacionais, o mainframe consolidou a ideia de centro de processamento de dados. Usuários acessavam sistemas por terminais, enquanto processamento, armazenamento, segurança e rotinas de backup eram controlados centralmente. A maturidade veio quando hardware, sistemas operacionais, linguagens como COBOL e práticas de operação formaram um ecossistema estável.

O declínio relativo ocorreu quando a economia de escala dos microprocessadores e servidores x86 passou a oferecer custo menor e maior flexibilidade. Mesmo assim, workloads transacionais de alto volume continuam usando mainframe porque a substituição completa pode ser cara, arriscada e desnecessária. (HEVNER; BERNDT, 2000)

Leitura executiva: confiabilidade, governança e continuidade são ativos. A substituição não deve ser confundida com abandono; a estratégia correta costuma ser integração por APIs e modernização seletiva.

4. Microprocessadores, PC/x86 e produtividade distribuída

Início em escala de mercado	1981–1990, com o IBM PC, clones compatíveis, MS-DOS, Windows e queda de preço dos microprocessadores.
Período de maturidade	1990–2005, quando desktops, notebooks, suítes de produtividade e servidores x86 tornaram-se padrão corporativo.
Declínio relativo	Declínio relativo como centro da inovação a partir de meados dos anos 2000, com web, mobilidade, SaaS e cloud deslocando o foco do dispositivo para o serviço.
Onda que absorveu o crescimento	Internet, aplicações web, SaaS, dispositivos móveis e plataformas cloud.



O microprocessador distribuindo capacidade computacional para postos de trabalho individuais

A segunda onda deslocou poder computacional para usuários e departamentos. O PC permitiu que analistas, gestores e áreas funcionais produzissem planilhas, documentos e relatórios sem depender integralmente do centro de processamento. O ecossistema x86

creceu porque combinou arquitetura aberta, fornecedores múltiplos, compatibilidade de software e economia de escala. (HEVNER; BERNDT, 2000)

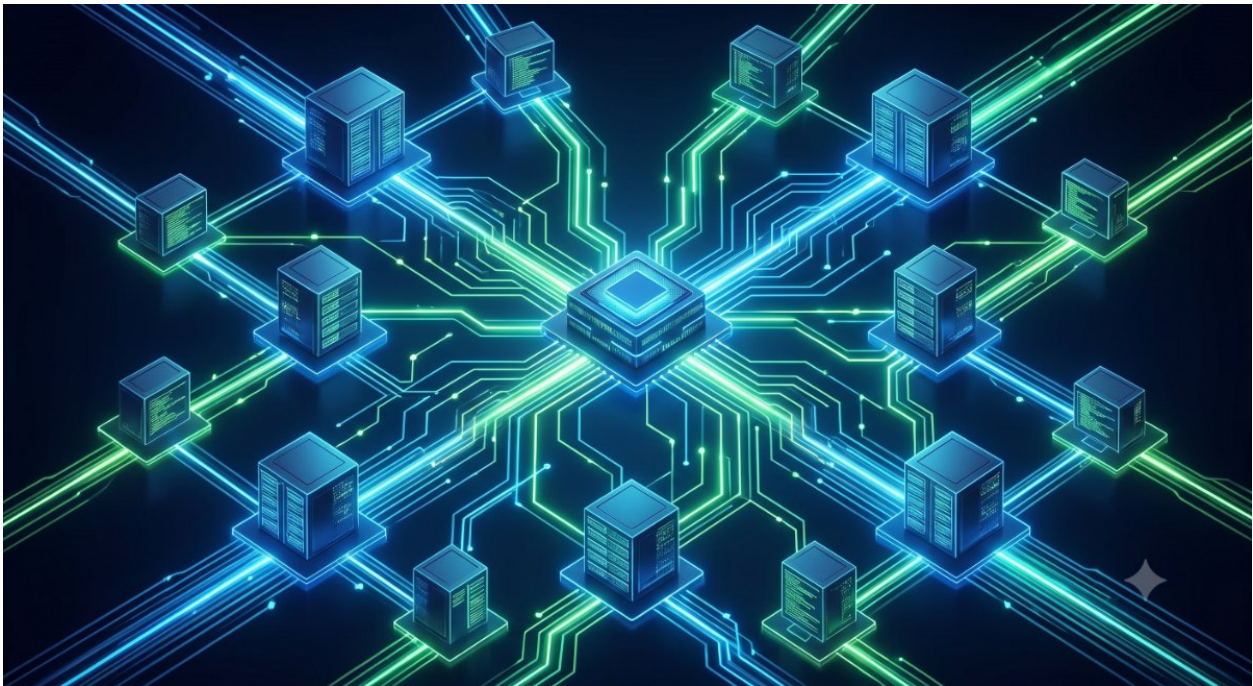
O impacto econômico foi enorme: computação, antes recurso central escasso, tornou-se ferramenta individual. Ao mesmo tempo, a descentralização criou novos problemas: arquivos duplicados, versões inconsistentes, segurança frágil, shadow IT e proliferação de servidores sem governança.

O declínio relativo não significou fim dos PCs. Eles permanecem essenciais como interfaces de trabalho. O que mudou foi o centro de valor: dados e lógica de negócio migraram para servidores remotos, e o endpoint tornou-se janela de acesso a serviços.

Leitura executiva: descentralização aumenta velocidade, mas exige governança. Essa tensão reaparece hoje com cloud self-service e IA generativa.

5. Redes locais, cliente-servidor e integração departamental

Início em escala de mercado	1985–1995, com Ethernet, servidores departamentais, bancos relacionais e aplicações cliente-servidor.
Período de maturidade	1995–2005, quando ERP, CRM, diretórios, e-mail corporativo e aplicações em rede tornaram-se padrão.
Declínio relativo	A partir dos anos 2000, quando a Internet, SOA, web applications e cloud reduziram a centralidade do cliente pesado e do servidor departamental isolado.
Onda que absorveu o crescimento	Aplicações web, arquitetura orientada a serviços, virtualização e cloud.



Redes locais conectando a produtividade individual em processos departamentais integrados

As redes locais conectaram a produtividade individual. Ethernet, protocolos de rede corporativa, servidores de arquivo e aplicações cliente-servidor permitiram que departamentos integrassem processos, sustentando ERPs, CRMs e sistemas de manufatura. (HEVNER; BERNDT, 2000)

O modelo cliente-servidor amadureceu porque respondia a uma necessidade prática: distribuir interface e processamento para o cliente, mantendo dados compartilhados no servidor. Seu limite apareceu na gestão de complexidade: aplicações cliente pesado exigiam instalação e atualização em muitos desktops, e servidores departamentais proliferavam.

O declínio relativo do cliente-servidor abriu espaço para aplicações web, middleware, APIs, SOA e, posteriormente, cloud. A arquitetura moderna ainda usa conceitos cliente-servidor, de forma mais abstrata.

Leitura executiva: integração cria valor, mas também aumenta dependências. Padrões de dados, APIs, identidade e observabilidade devem ser planejados desde o início para evitar silos futuros.

6. Internet comercial, web e plataformas digitais

Início em escala de mercado	1993–2000, com navegadores, provedores comerciais, websites, e-commerce e hospedagem web.
Período de maturidade	2000–2015, quando comércio eletrônico, busca, redes sociais, mídia digital e serviços online tornaram-se infraestrutura econômica.
Declínio relativo	A Internet não declina como base; o que declina é o modelo web monolítico e pouco escalável, substituído por cloud, mobile, APIs e plataformas distribuídas.
Onda que absorveu o crescimento	Cloud pública, mobile, microsserviços, plataformas digitais e hyperscale.



A internet global conectando mercados, dados e ecossistemas digitais em escala planetária

A Internet transformou a TI em infraestrutura de mercado. Protocolos abertos como TCP/IP, DNS, HTTP e SMTP permitiram interoperabilidade global. A web converteu sistemas internos em canais de relacionamento, venda, conteúdo e serviço. (LEINER et al., 1997; TOWNES, 2012)

A maturidade da Internet ocorreu quando empresas entenderam que presença digital não era apenas marketing. A disponibilidade passou a impactar receita: uma falha de servidor deixou de afetar apenas usuários internos e passou a interromper clientes em escala global.

A Internet também mudou a engenharia de infraestrutura, tornando essenciais balanceamento de carga, DNS resiliente, firewalls, CDN, monitoramento, criptografia e data centers conectados a múltiplas operadoras.

Leitura executiva: conectividade transforma sistemas em canais de receita e reputação. A Internet mostrou que padrões abertos tendem a vencer ecossistemas fechados quando o objetivo é escala global.

7. Virtualização: ponte entre o servidor físico e a cloud

Início em escala de mercado	1999–2008, com virtualização x86 empresarial e consolidação de servidores.
Período de maturidade	2008–2015, quando máquinas virtuais, clusters, templates, snapshots e automação tornaram-se práticas maduras.
Declínio relativo	Declínio relativo da VM como unidade principal de inovação a partir da cloud nativa,

Onda que absorveu o crescimento

containers, Kubernetes e serverless; permanece amplamente usada.
Cloud pública, containers, infraestrutura como código, PaaS e serverless.



A virtualização abstraindo o hardware físico em ambientes replicáveis e consolidados

A virtualização foi a ponte entre data centers tradicionais e cloud. Ela abstrai o hardware físico, permitindo executar múltiplas máquinas virtuais em um único servidor, aumentando utilização e simplificando recuperação. (OPPITZ; TOMSU, 2018)

A escala de mercado veio quando servidores x86 se tornaram suficientemente potentes e ferramentas de virtualização amadureceram. A virtualização introduziu conceitos que depois seriam naturais na cloud: pool de recursos, migração de cargas, provisionamento rápido e templates.

O declínio relativo ocorreu quando a unidade de inovação passou da máquina virtual para plataformas mais abstratas: containers, Kubernetes e serverless. Mesmo assim, VMs continuam fundamentais em cloud e ambientes híbridos.

Leitura executiva: virtualização não é apenas eficiência técnica, é mudança de mentalidade. Ela transforma ativos físicos em capacidade compartilhada, lógica que a cloud levou adiante.

8. Cloud computing: surgimento, escala, maturidade e transformação financeira

Início em escala de mercado	2006–2015, com serviços comerciais de infraestrutura sob demanda, regiões globais e adoção por startups e empresas.
Maturidade	2015–presente, com IaaS, PaaS, SaaS, cloud híbrida, governança, segurança, FinOps e serviços de IA.
Declínio relativo	Não há declínio estrutural. Há maturação, especialização e correção de excessos, com repatriação seletiva de workloads quando custo, latência ou controle justificam.
Nova onda complementar	Multi-cloud, hybrid cloud, local zones, edge, GPU cloud e AI factories.



A cloud transformando infraestrutura física em capacidade computacional sob demanda

A cloud resultou da convergência de virtualização, automação, redes IP, armazenamento distribuído, data centers em escala, APIs e modelos comerciais mensuráveis. A definição do NIST caracteriza cloud computing por cinco propriedades: autosserviço sob demanda, amplo acesso à rede, pooling de recursos, elasticidade rápida e serviço mensurável. (MELL; GRANCE, 2011; ARMBRUST et al., 2010) (MELL; GRANCE, 2011)

A escala de mercado começou quando provedores passaram a vender capacidade computacional por API, atraindo startups que precisavam crescer sem comprar infraestrutura antecipadamente. A maturidade veio quando cloud passou a oferecer catálogos amplos de serviços: bancos gerenciados, analytics, machine learning e automação.

8.1 Cloud como modelo financeiro

Cloud altera a lógica de investimento, deslocando CAPEX para OPEX e permitindo alinhar custo à atividade. Essa realidade criou a disciplina de FinOps, que integra engenharia, finanças e negócio para maximizar valor por unidade de consumo.

8.2 Cloud como modelo operacional

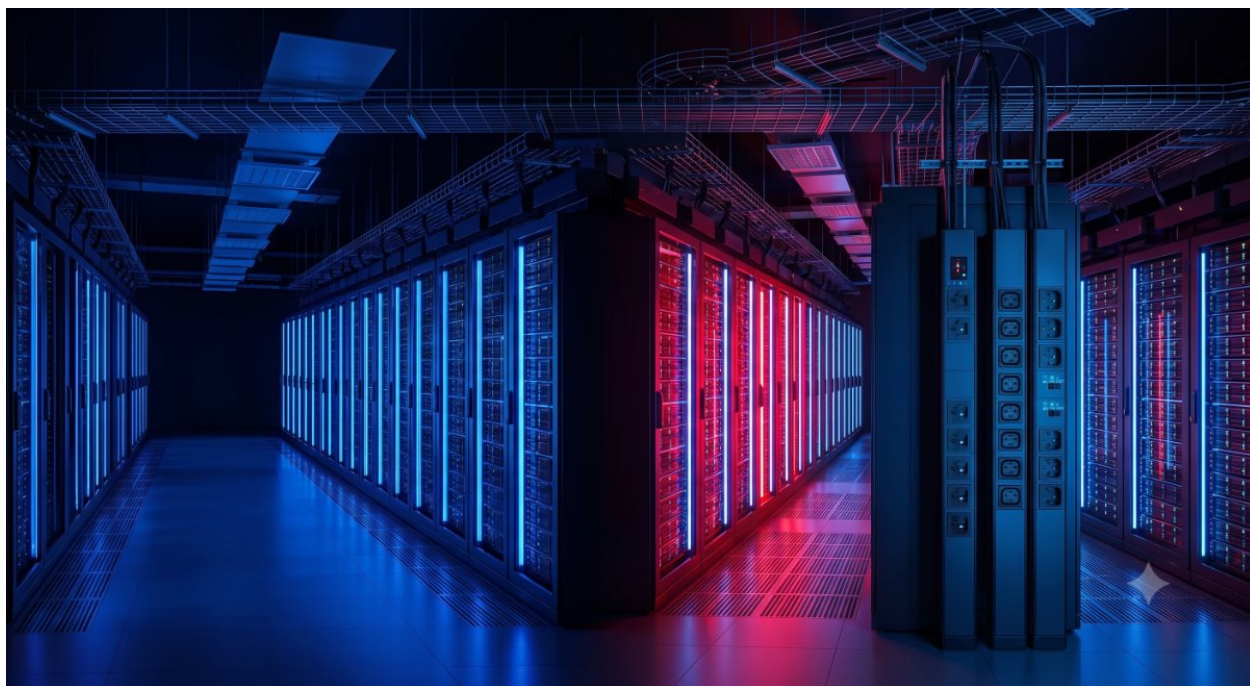
Operacionalmente, cloud substitui processos manuais por automação. Infraestrutura como código e pipelines reduzem tempo de entrega, mas mudam a natureza do risco: um erro de configuração pode criar exposição em minutos. A responsabilidade é compartilhada entre provedor e cliente.

8.3 Cloud como plataforma de inovação

A cloud acelerou a criação de produtos digitais ao reduzir a distância entre ideia e ambiente de produção. O limite da cloud aparece quando custo, soberania, latência ou controle exigem alternativas como colocation ou infraestrutura dedicada. A maturidade não é "tudo em cloud", mas uma arquitetura híbrida e governada. (ARMBRUST et al., 2010)

9. O que é um data center: explicação simples, técnica e executiva

Um data center é um ambiente projetado para manter computadores funcionando com energia, refrigeração, conectividade, segurança e operação contínua. A analogia mais clara é a de uma fábrica de serviços digitais: a matéria-prima é energia e dados, as máquinas são servidores, aceleradores e equipamentos de rede, o produto final é disponibilidade, processamento, armazenamento e capacidade de executar aplicações. (FOGARTY; FLUCKER, 2023; GENG, 2021)



Corredores frios e quentes: a disciplina física que sustenta a disponibilidade do data center

9.1 Os componentes físicos: sala de TI, racks e equipamentos

A sala de TI é o espaço onde ficam racks com servidores, storages, switches e equipamentos especializados. Em IA, um único rack pode exigir potência muito maior e gerar calor concentrado. A maturidade de um data center começa pela disciplina física: inventário, identificação de cabos, controle de acesso e documentação.

9.2 Energia: o sistema circulatório do data center

Energia é o primeiro requisito crítico. A UPS sustenta a carga em oscilações e dá tempo para geradores assumirem em interrupções prolongadas. Redundância energética é expressa por desenhos como N, N+1, 2N e 2N+1. A IA torna energia ainda mais estratégica: o gargalo de um projeto pode deixar de ser o prédio e passar a ser a disponibilidade de megawatts e transformadores. (GENG, 2021)

9.3 Refrigeração: retirar calor para manter computação

Computadores transformam energia elétrica em trabalho computacional e calor. Data centers usam corredores frios e quentes, CRAC, CRAH, chillers e torres de resfriamento. Com IA, a densidade aumenta e soluções líquidas ganham relevância: refrigeração líquida direta no chip, rear-door heat exchangers e imersão. O PUE compara energia total consumida pelo data center com energia consumida pelos equipamentos de TI. (FOGARTY; FLUCKER, 2023; GENG, 2021)

9.4 Conectividade: transformar prédio em ecossistema

Um data center sem conectividade é uma sala de computadores isolada. O valor cresce quando há múltiplas operadoras, pontos de troca de tráfego, cloud on-ramps e interconexão com parceiros. Localização importa: proximidade de energia, cabos submarinos e ecossistemas financeiros ou industriais gera valor de rede. (GENG, 2021)

9.5 Segurança, compliance e operação 24x7

A segurança física protege o ativo material: controle de acesso, biometria, câmeras e procedimentos de escolta. A operação 24x7 transforma o prédio em serviço, com equipes de NOC, facilities e gestão de mudança coordenando atividades contínuas.

9.6 Da infraestrutura ao produto: modelos comerciais

O data center pode ser monetizado em diferentes camadas: corporativo próprio, colocation, bare metal, cloud IaaS, PaaS/SaaS e GPU-as-a-Service. A pergunta estratégica não é "qual modelo é melhor?", mas "qual combinação de modelos otimiza cada carga?".

10. Colocation, interconexão e local zones

Início em escala de mercado	Colocation moderno: 1998–2010; expansão com Internet, telecom e terceirização de data centers.
Maturidade	2010–presente, com interconexão, cloud on-ramps, campi hyperscale, alta densidade e presença regional.
Declínio relativo	Não há declínio; há reposicionamento. O modelo deixa de vender apenas espaço e passa a vender ecossistema, energia, conectividade e densidade.
Nova onda complementar	Hybrid cloud, local zones, edge, soberania de dados e IA.



Fibra óptica e interconexão: o tecido físico que conecta operadoras, nuvens e parceiros

O colocation cresceu porque muitas empresas não queriam operar data centers próprios, mas também não podiam migrar tudo para cloud pública. A maturidade do colocation veio com a interconexão: um data center neutro que reúne operadoras, nuvens e empresas cria efeitos de rede. (GENG, 2021)

Local zones e edge respondem a limites da centralização. Capacidade computacional em áreas metropolitanas ou industriais permite aplicações de varejo, jogos, streaming, manufatura, saúde e inferência de IA. Edge não substitui hyperscale; complementa.

A maturidade executiva está em entender que colocation é opção estratégica, não apenas terceirização de sala. Com a pressão por energia e GPUs, operadores capazes de oferecer megawatts, refrigeração avançada e conectividade privada tornam-se parte central da cadeia de IA.

11. Hyperscale e computação em escala de armazém

Início em escala de mercado	2010–2018, com expansão global de cloud pública, redes sociais, busca, streaming e grandes plataformas digitais.
Maturidade	2018–presente, com regiões globais, hardware customizado, automação extrema, compras de energia e software distribuído.
Declínio relativo	Não há declínio; há pressão por energia, regulação, cadeia de suprimentos e especialização para IA.
Nova onda complementar	AI factories, GPU clusters, cloud soberana, edge e infraestrutura acelerada.



Automação em escala de armazém: milhares de servidores coordenados como um único computador

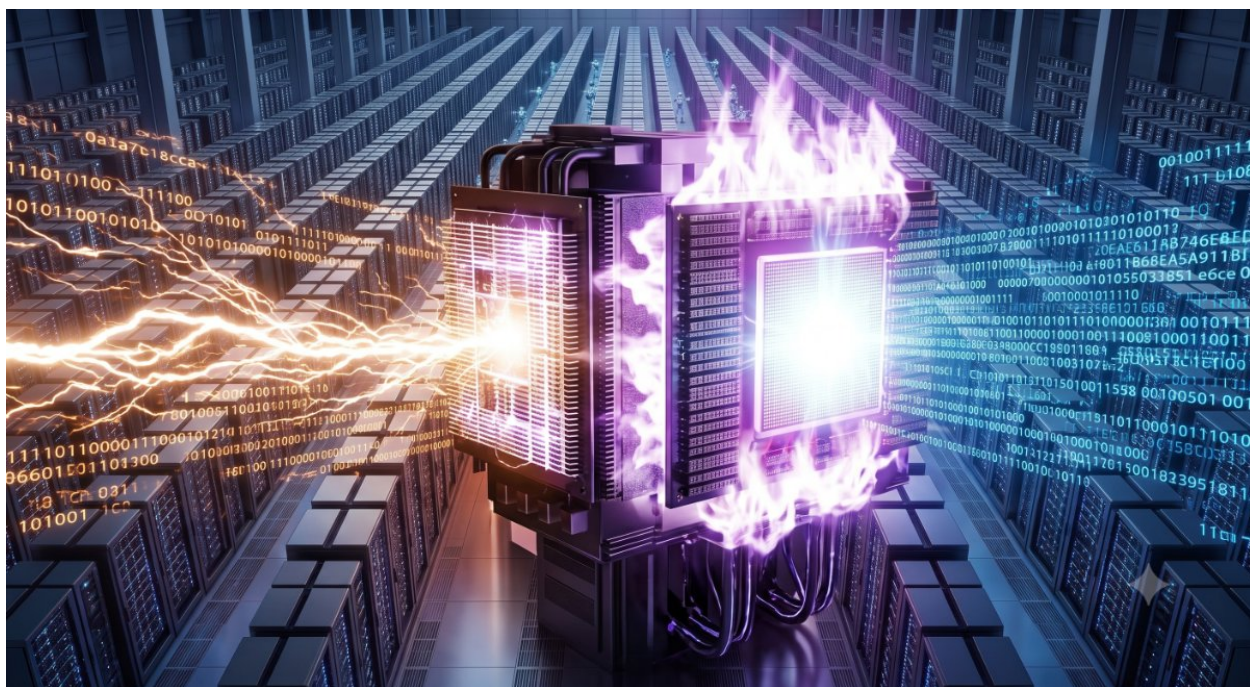
Hyperscale representa a industrialização extrema da infraestrutura digital. Barroso, Hölzle e Ranganathan (2018) descrevem o data center de escala de armazém como um computador gigantesco, formado por milhares de servidores, redes internas e software capaz de coordenar falhas e balancear carga. (BARROSO; HÖLZLE; RANGANATHAN, 2018)

Hyperscalers como AWS, Microsoft, Google, Meta, Apple, Alibaba e Tencent operam com escala suficiente para influenciar mercados de servidores, semicondutores, cabos, imóveis e energia. A IA pressiona o hyperscale porque treinamento de modelos exige clusters densos, redes de baixa latência e armazenamento de alto desempenho.

A lição executiva é que escala física e escala de software são inseparáveis. Países que desejam competir na economia digital precisam considerar energia, conectividade, regulação e atração de data centers como política industrial.

12. IA generativa e AI factories

Início em escala de mercado	IA generativa: 2022–2025, com modelos fundacionais, copilotos e APIs comerciais. AI factories: 2023–2026, com clusters dedicados e infraestrutura acelerada.
Maturidade	Em rápida evolução. Maturidade técnica e econômica ainda em consolidação, especialmente em governança, custo por inferência e retorno de negócio.
Declínio relativo	Ainda não há declínio. O ciclo atual está em expansão, com tendência de consolidação de fornecedores e otimização de modelos.
Nova onda complementar	Agentes autônomos, IA multimodal, inferência distribuída, modelos menores



A AI factory: energia, chips e dados convertidos em inteligência operacional em escala industrial

Uma AI factory transforma dados, energia, chips, modelos, pipelines e talento técnico em inteligência operacional. Assim como uma fábrica industrial precisa de máquinas, energia, insumos e controle de qualidade, uma AI factory precisa de GPUs ou TPUs, redes de alta velocidade, armazenamento, MLOps, segurança e governança. (REED; GANNON; DONGARRA, 2022; RANGANATHAN; HÖLZLE, 2024)

12.1 AI factories e mercado global de energia

A demanda de IA altera a relação entre tecnologia e energia. Data centers tornam-se grandes compradores de potência firme, contratos renováveis e subestações. O debate deixa de ser apenas preço por quilowatt-hora e passa a incluir tempo de conexão à rede, capacidade de expansão e intensidade de carbono. (MASANET et al., 2020; SHEHABI et al., 2018) (INTERNATIONAL ENERGY AGENCY, 2024)

12.2 AI factories e modelos de negócio

A AI factory pode ser propriedade de hyperscalers, operadores especializados, governos ou consórcios, com modelos que vão de GPU-as-a-Service a parcerias público-privadas. O maior risco executivo é confundir uso de IA com vantagem competitiva:

consumir uma API pública pode aumentar produtividade, mas não gerar diferenciação sustentável.

13. Quadro integrado: evolução, maturidade e declínio relativo

Onda	Proposta original	Maturidade atual	Declínio relativo / limite	Agenda executiva
Mainframe	Centralizar transações e governança	Alta em workloads críticos	Custo, rigidez e interfaces fechadas	Modernizar por APIs; evitar migração sem caso econômico
PC/x86	Democratizar produtividade	Alta como endpoint	Fragmentação e segurança	Gerir endpoint como acesso a serviços
Cliente-servidor	Integrar departamentos	Média/alta em legados	Cliente pesado e silos	Modernizar para web, APIs e cloud
Internet/Web	Conectar mercados	Base estrutural madura	Monólitos e hospedagem simples	Escalar via cloud, CDN e observabilidade
Virtualização	Consolidar servidores	Alta, mas madura	VM sprawl e abstração limitada	Automatizar e complementar com containers
Cloud	Elasticidade e serviços	Alta e crescente	Custo sem governança e lock-in	FinOps, arquitetura híbrida e segurança
Colocation	Resiliência e interconexão	Alta	Commodity se vender só espaço	Diferenciar por energia, ecossistema e densidade
Hyperscale	Industrializar infraestrutura	Alta	Pressão física e regulatória	Energia, sustentabilidade e IA acelerada
AI Factory	Produzir inteligência	Emergente	Ainda indefinido	Governança, unit economics e energia

O padrão é recorrente: cada onda resolve um problema e cria outro. O mainframe resolveu confiabilidade, mas criou centralização rígida. O PC resolveu autonomia, mas criou fragmentação. A cloud trouxe elasticidade, mas introduziu governança de consumo. O hyperscale reduziu custo unitário, mas trouxe restrições físicas de energia. A IA promete inteligência, mas exige capacidade computacional, dados, governança e sustentabilidade.

14. Implicações para estratégia corporativa

A evolução do mainframe às AI factories muda a agenda da alta gestão. Infraestrutura digital deixa de ser assunto apenas de tecnologia e passa a compor estratégia competitiva. Um banco pode preservar mainframe para core transacional e usar cloud para analytics; uma indústria pode usar edge e IA no chão de fábrica. (HENNESSY; PATTERSON, 2019; RANGANATHAN; HÖLZLE, 2024)

A primeira implicação é portfólio: não existe arquitetura universal. A segunda é custo: cloud e IA transformam infraestrutura em consumo variável, exigindo FinOps e

unit economics. A terceira é risco: dependência de um provedor, região ou fonte de energia precisa ser conhecida e mitigada.

A quarta implicação é talento: operar a nova infraestrutura requer competências em cloud, segurança, dados, MLOps e sustentabilidade. A quinta é sustentabilidade, exigindo transparência sobre energia, carbono e água. A sexta é soberania de dados, modelos e chaves criptográficas.

A recomendação executiva é criar uma governança integrada de infraestrutura digital, reunindo tecnologia, finanças, risco, jurídico, sustentabilidade e áreas de negócio para decidir onde cada carga deve rodar, com qual custo, risco e trajetória de evolução

15. Roteiro de maturidade para organizações

Uma organização que deseja evoluir do legado à IA deve seguir um roteiro pragmático: inventariar aplicações, dados, dependências e custos; classificar workloads por criticidade e sensibilidade; definir arquitetura alvo híbrida; modernizar integração via APIs; estabelecer FinOps e governança de cloud; criar governança de dados e IA; planejar infraestrutura de IA (APIs públicas, modelos privados ou capacidade dedicada); revisar energia e sustentabilidade; treinar pessoas; e medir retorno com métricas que incluam acurácia, custo por inferência e impacto operacional.

16. Conclusão

A trajetória do mainframe às AI factories mostra que a história da TI não é apenas a história de máquinas mais rápidas. É a história de modelos econômicos e operacionais. O mainframe centralizou confiança. O PC distribuiu produtividade. As redes locais integraram departamentos. A Internet conectou mercados. A virtualização consolidou recursos. A cloud transformou infraestrutura em serviço. O colocation criou ecossistemas neutros de interconexão. O hyperscale industrializou software e hardware em escala global. A IA transforma capacidade computacional em inteligência produtiva.

Cada onda teve data de escala, maturidade e declínio relativo. Nenhuma desapareceu completamente; cada uma encontrou novo papel. O futuro não será uma substituição simples de tudo por IA. Será a combinação de legados confiáveis, cloud, edge, hyperscale, dados governados, energia disponível e AI factories especializadas.



O relatório executivo "Da Era do Mainframe às AI Factories": a materialidade da economia digital

O ponto mais importante é que data centers se tornaram infraestrutura crítica da economia. Empresas e países que compreenderem essa materialidade poderão capturar valor. Aqueles que tratarem infraestrutura como detalhe invisível enfrentarão gargalos, dependência e perda de competitividade.

A decisão executiva contemporânea é decidir onde e como produzir capacidade digital. O diferencial não estará em seguir uma moda tecnológica, mas em alinhar cada onda ao valor de negócio, ao risco, ao custo, à energia e à maturidade organizacional.

REFERÊNCIAS

- ARMBRUST, M.; FOX, A.; GRIFFITH, R.; JOSEPH, A. D.; KATZ, R.; KONWINSKI, A.; LEE, G.; PATTERSON, D.; RABKIN, A.; STOICA, I.; ZAHARIA, M. A view of cloud computing. *Communications of the ACM*, v. 53, n. 4, p. 50–58, 2010.
- BARROSO, L. A.; HÖLZLE, U.; RANGANATHAN, P. *The Datacenter as a Computer: Designing Warehouse-Scale Machines*. Morgan & Claypool, 2018.
- EVANS, B. O. System/360: A retrospective view. *Annals of the History of Computing*, v. 8, n. 2, p. 155–179, 1986.
- FOGARTY, V.; FLUCKER, S. *Data Centre Essentials: Design, Construction, and Operation of Data Centres for the Non-expert*. BCS Learning & Development, 2023.
- GENG, H. *Data Center Handbook: Plan, Design, Build, and Operations of a Smart Data Center*. Wiley, 2021.
- HENNESSY, J. L.; PATTERSON, D. A. A new golden age for computer architecture. *Communications of the ACM*, v. 62, n. 2, p. 48–60, 2019.
- HEVNER, A. R.; BERNDT, D. J. Eras of business computing. *Advances in Computers*, v. 51, p. 1–84, 2000.
- INTERNATIONAL ENERGY AGENCY. *Electricity and data centres analysis*. 2024.
- LEINER, B. M. et al. The past and future history of the Internet. *Communications of the ACM*, v. 40, n. 2, p. 102–108, 1997.
- MASANET, E.; SHEHABI, A.; LEI, N.; SMITH, S.; KOOMEY, J. Recalibrating global data center energy-use estimates. *Science*, v. 367, n. 6481, p. 984–986, 2020.
- MELL, P.; GRANCE, T. *The NIST Definition of Cloud Computing*. NIST Special Publication 800-145. Gaithersburg: National Institute of Standards and Technology, 2011.
- OPPITZ, M.; TOMSU, P. *Inventing the Cloud Century*. Springer, 2018.
- PILZ, K.; HEIM, L. Compute at scale: A broad investigation into the data center industry. *arXiv:2311.15377*, 2023.
- PUGH, E. W.; JOHNSON, L. R.; PALMER, J. H. *IBM's 360 and Early 370 Systems*. MIT Press, 1991.
- RANGANATHAN, P.; HÖLZLE, U. Twenty five years of warehouse-scale computing. *IEEE Micro*, 2024.
- REED, D.; GANNON, D.; DONGARRA, J. Reinventing high performance computing. *arXiv:2203.02544*, 2022.
- SHEHABI, A.; SMITH, S. J.; MASANET, E.; KOOMEY, J. Data center growth in the United States. *Environmental Research Letters*, v. 13, n. 12, 124030, 2018.
- TOWNES, M. The spread of TCP/IP: How the Internet became the Internet. *Millennium*, v. 41, n. 1, p. 43–64, 2012.
- UPTIME INSTITUTE. *Annual data center survey and resiliency reports*. 2024.