大模型系列 – 大模型算法

打破 Transformer

ZOMI

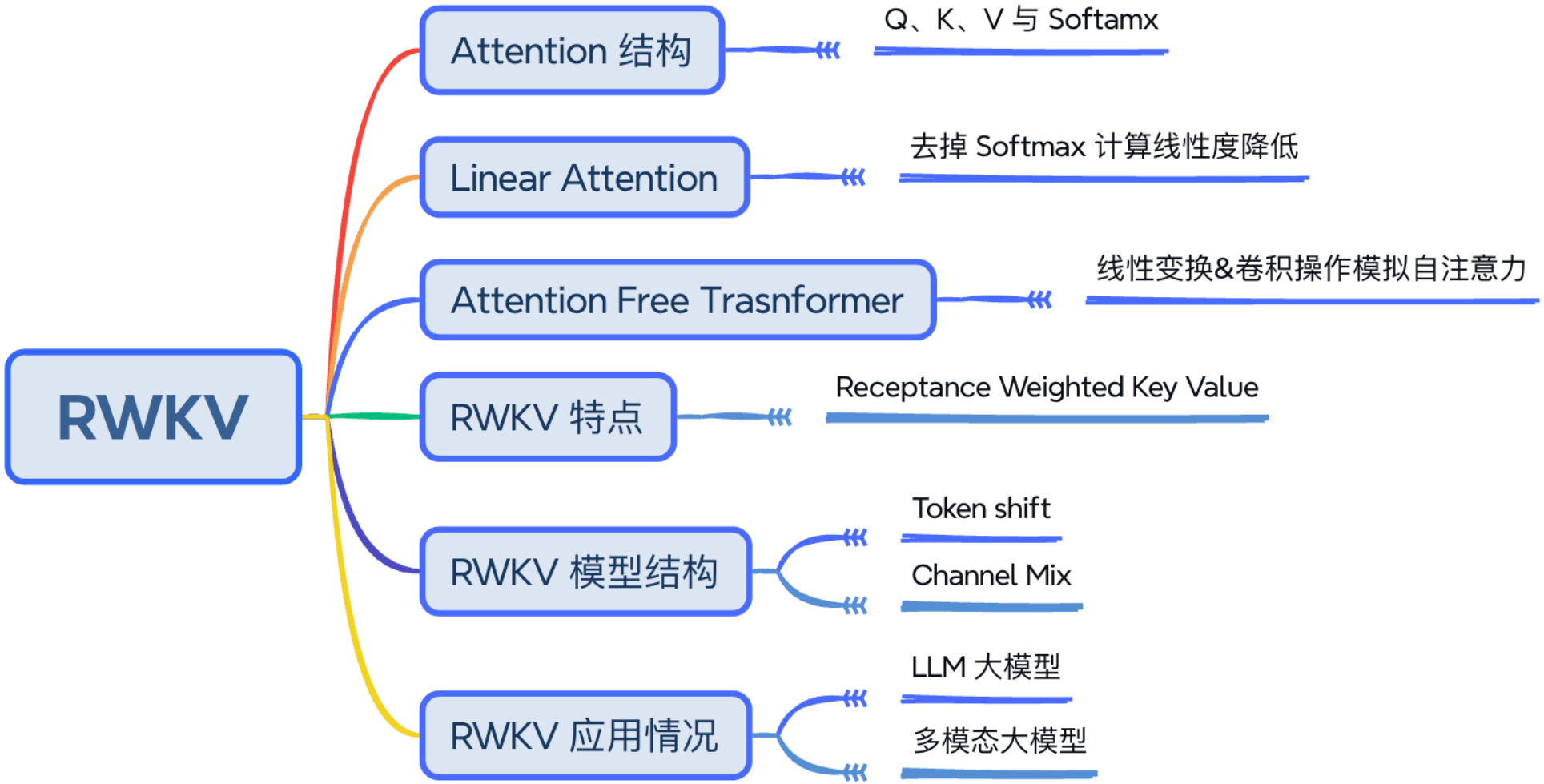
关于本内容

1. Transformer 结构回顾与挑战
2. RWKV 新结构
3. Mamba 新结构
4. 未来打破 Transformer 路在何方

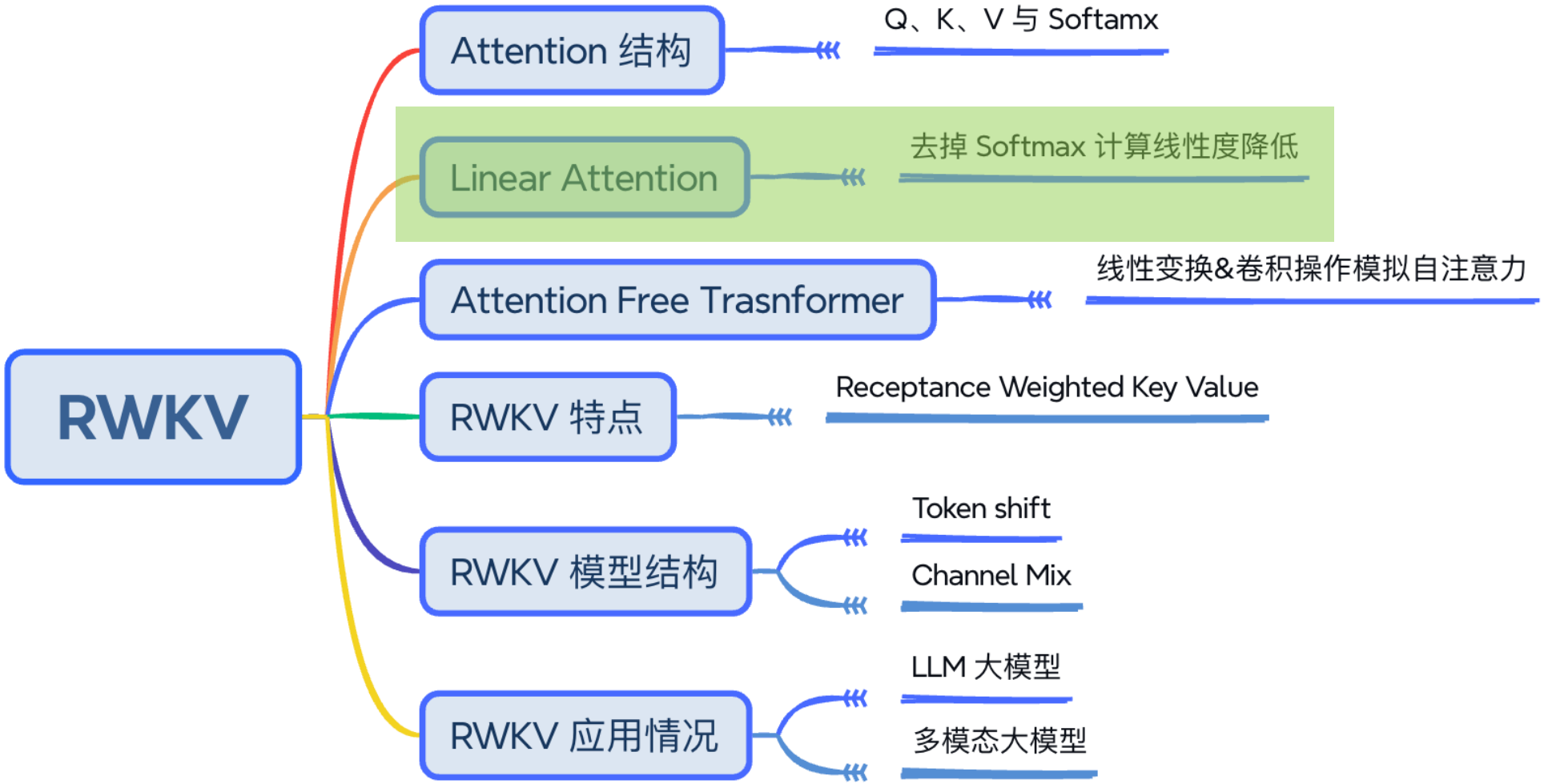
01

RWKV

RWKV


RWKV
Attention 结构
Q、K、V 与 Softamx
Linear Attention
去掉 Softmax 计算线性度降低
Attention Free Trasnformer
线性变换&卷积操作模拟自注意力
RWKV 特点
Receptance Weighted Key Value
RWKV 模型结构
Token shift
Channel Mix
RWKV 应用情况
LLM 大模型
多模态大模型

RWKV


RWKV
Attention 结构
Q、K、V 与 Softamx
Linear Attention
去掉 Softmax 计算线性度降低
Attention Free Trasnformer
线性变换&卷积操作模拟自注意力
RWKV 特点
Receptance Weighted Key Value
RWKV 模型结构
Token shift
Channel Mix
RWKV 应用情况
LLM 大模型
多模态大模型

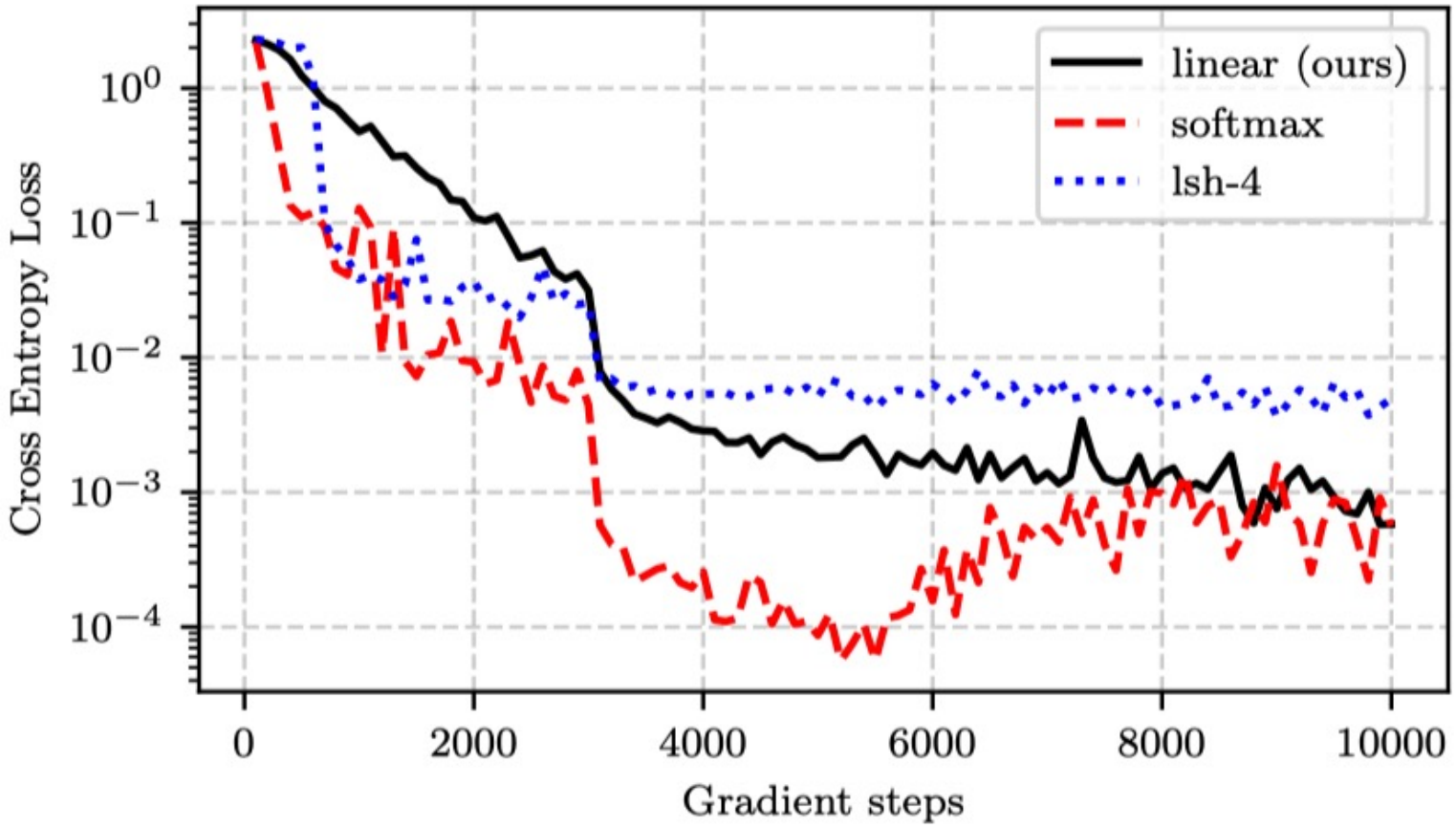
Linear Attention

- Transformers are rnns: Fast autoregressive transformers with linear attention.

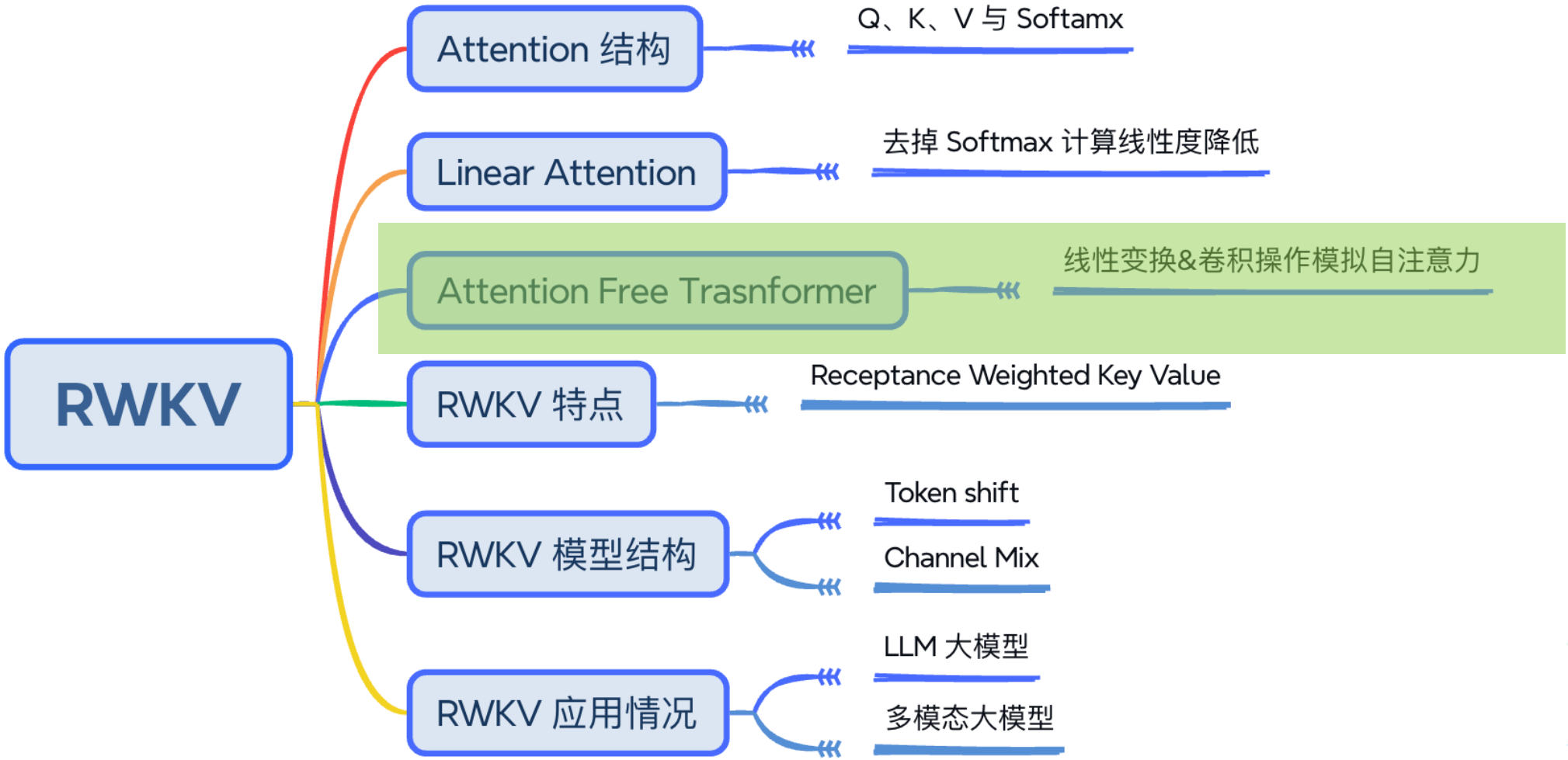
Algorithm 1 Linear transformers with causal masking

function forward $(\phi(Q), \phi(K), V)$**:**
  $V' \leftarrow 0, S \leftarrow 0$
  for $i = 1, \ldots, N$ **do**
    $S \leftarrow S + \phi(K_i) V_i^T$ equation 10
    $\bar{V}_i \leftarrow \phi(Q_i) S$
  end
  return $\bar{V}$
end

function backward $(\phi(Q), \phi(K), V, G)$**:**
  /* G is the gradient of the loss with respect to the output of forward */
  $S \leftarrow 0, \nabla_{\phi(Q)}\mathcal{L} \leftarrow 0$
  for $i = 1, \ldots, N$ **do**
    $S \leftarrow S + \phi(K_i) V_i^T$
    $\nabla_{\phi(Q_i)}\mathcal{L} \leftarrow G_i S^T$ equation 13
  end
  $S \leftarrow 0, \nabla_{\phi(K)}\mathcal{L} \leftarrow 0, \nabla_V \mathcal{L} \leftarrow 0$
  for $i = N, \ldots, 1$ **do**
    $S \leftarrow S + \phi(Q_i) G_i^T$
    $\nabla_{V_i}\mathcal{L} \leftarrow S^T \phi(K_i)$ equation 15
    $\nabla_{\phi(K_i)}\mathcal{L} \leftarrow S V_i$ equation 14
  end
  return $\nabla_{\phi(Q)}\mathcal{L}, \nabla_{\phi(K)}\mathcal{L}, \nabla_V \mathcal{L}$
end



RWKV


RWKV
Attention 结构
Q、K、V 与 Softamx
Linear Attention
去掉 Softmax 计算线性度降低
Attention Free Trasnformer
线性变换&卷积操作模拟自注意力
RWKV 特点
Receptance Weighted Key Value
RWKV 模型结构
Token shift
Channel Mix
RWKV 应用情况
LLM 大模型
多模态大模型

AFT

- Paper： An attention free transformer.

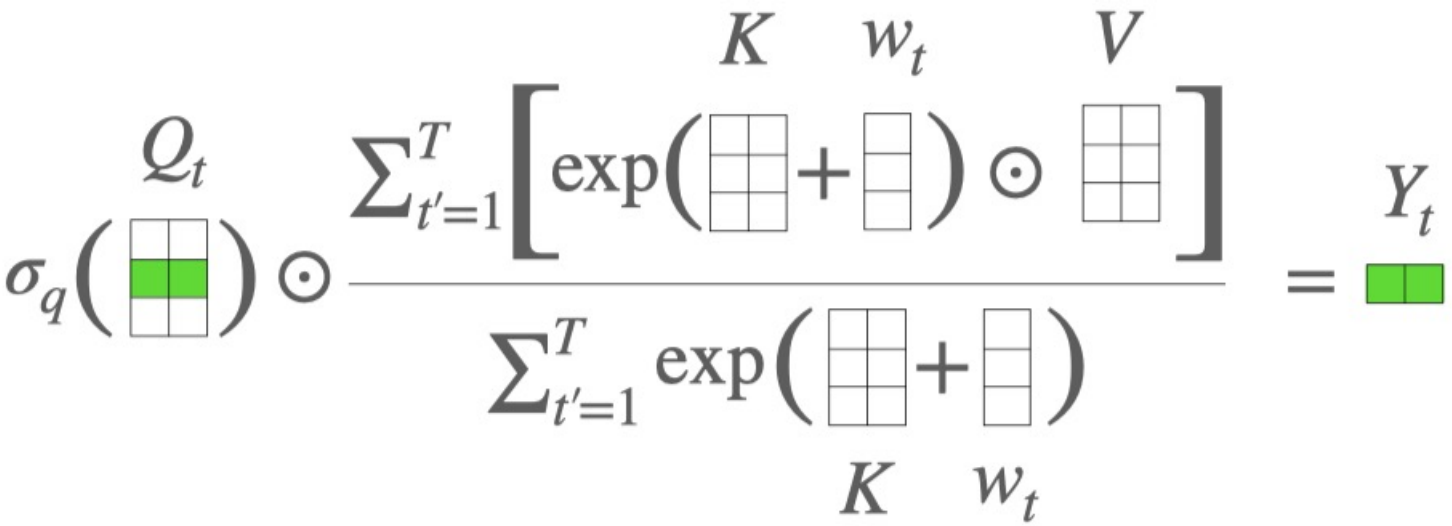
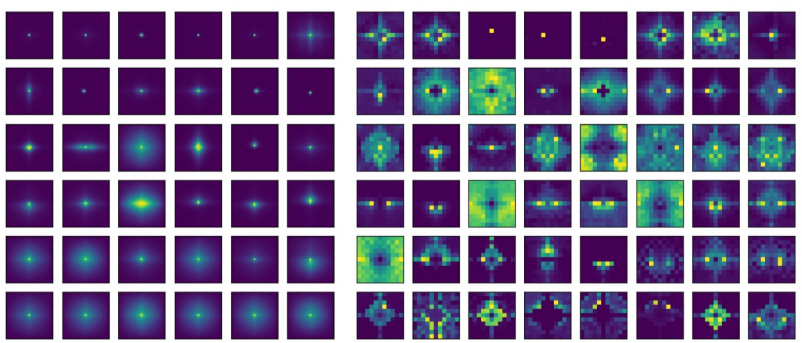


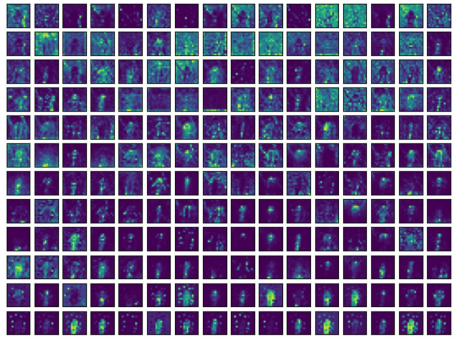
Figure 2: An illustration of AFT defined in Equation 2, with $T = 3, d = 2$.

AFT

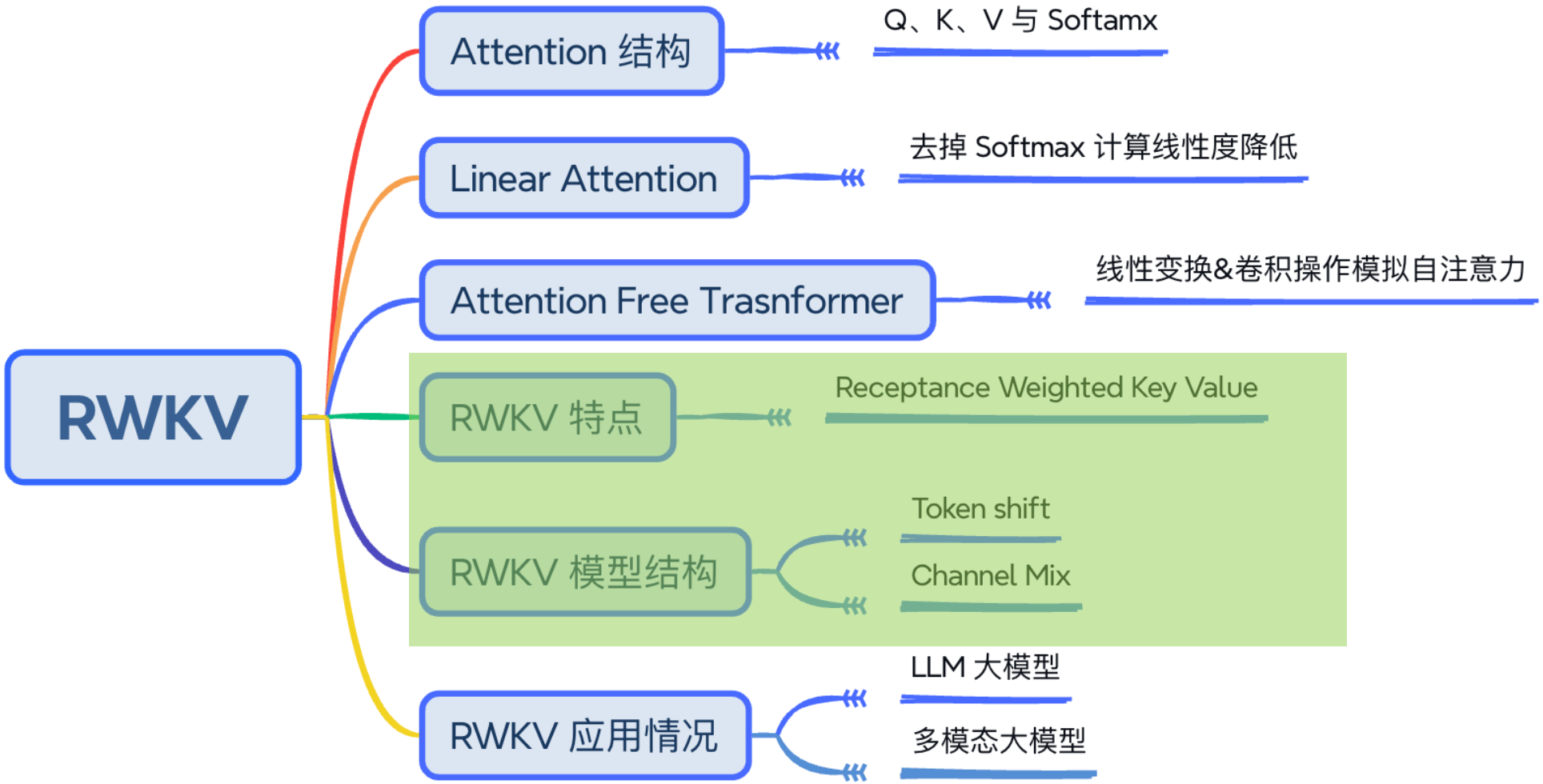
- Paper： An attention free transformer.

Model	Time	Space
Transformer	$O(T^2d)$	$O(T^2 + Td)$
Reformer	$O(T \log Td)$	$O(T \log T + Td)$
Linear Transformer	$O(Td^2)$	$O(Td + d^2)$
Performer	$O(Td^2 \log d)$	$O(Td \log d + d^2 \log d)$
AFT-simple	$O(\mathbf{Td})$	$O(\mathbf{Td})$
AFT-full	$O(T^2d)$	$O(\mathbf{Td})$
AFT-local (AFT-conv)	$O(Tsd),\ s < T$	$O(\mathbf{Td})$



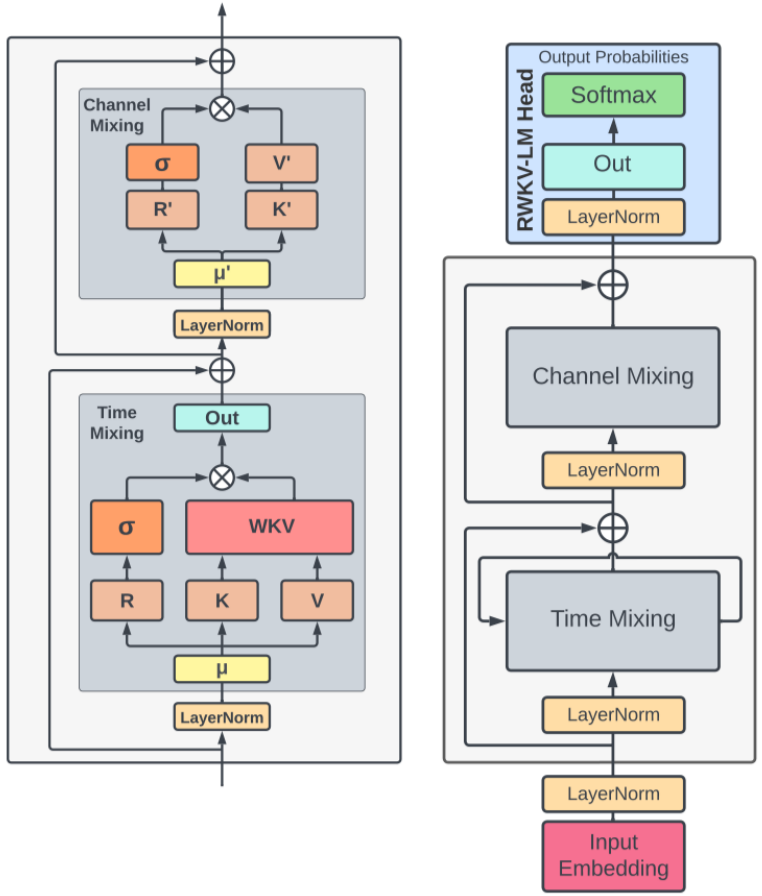


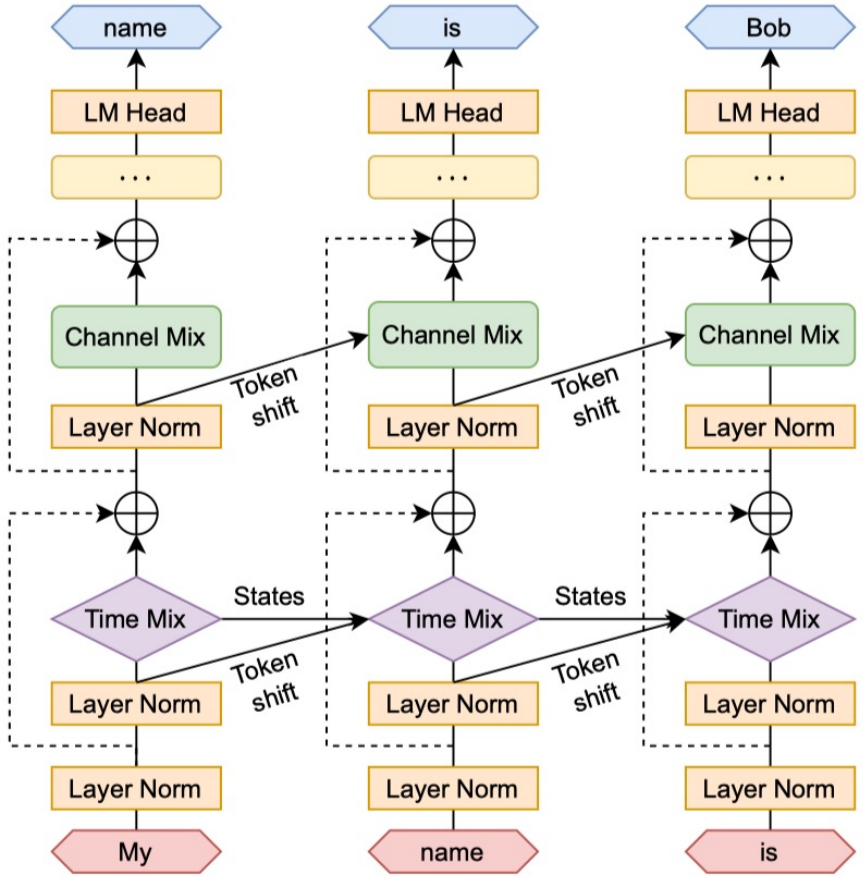
RWKV


RWKV
Attention 结构
Q、K、V 与 Softamx
Linear Attention
去掉 Softmax 计算线性度降低
Attention Free Trasnformer
线性变换&卷积操作模拟自注意力
RWKV 特点
Receptance Weighted Key Value
RWKV 模型结构
Token shift
Channel Mix
RWKV 应用情况
LLM 大模型
多模态大模型

RWKV

- RWKV: Reinventing RNNs for the Transformer Era



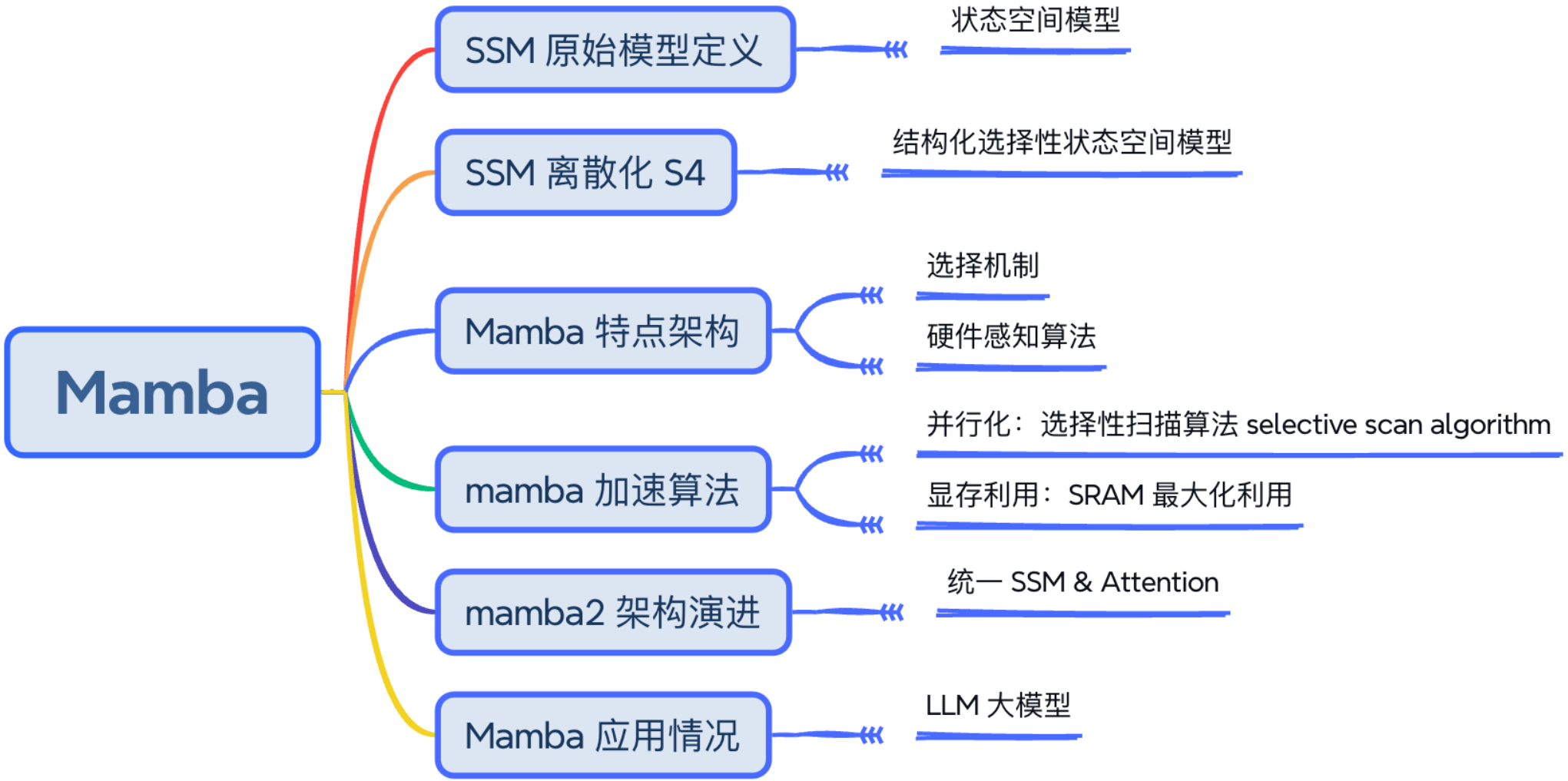


Reference 参考&引用

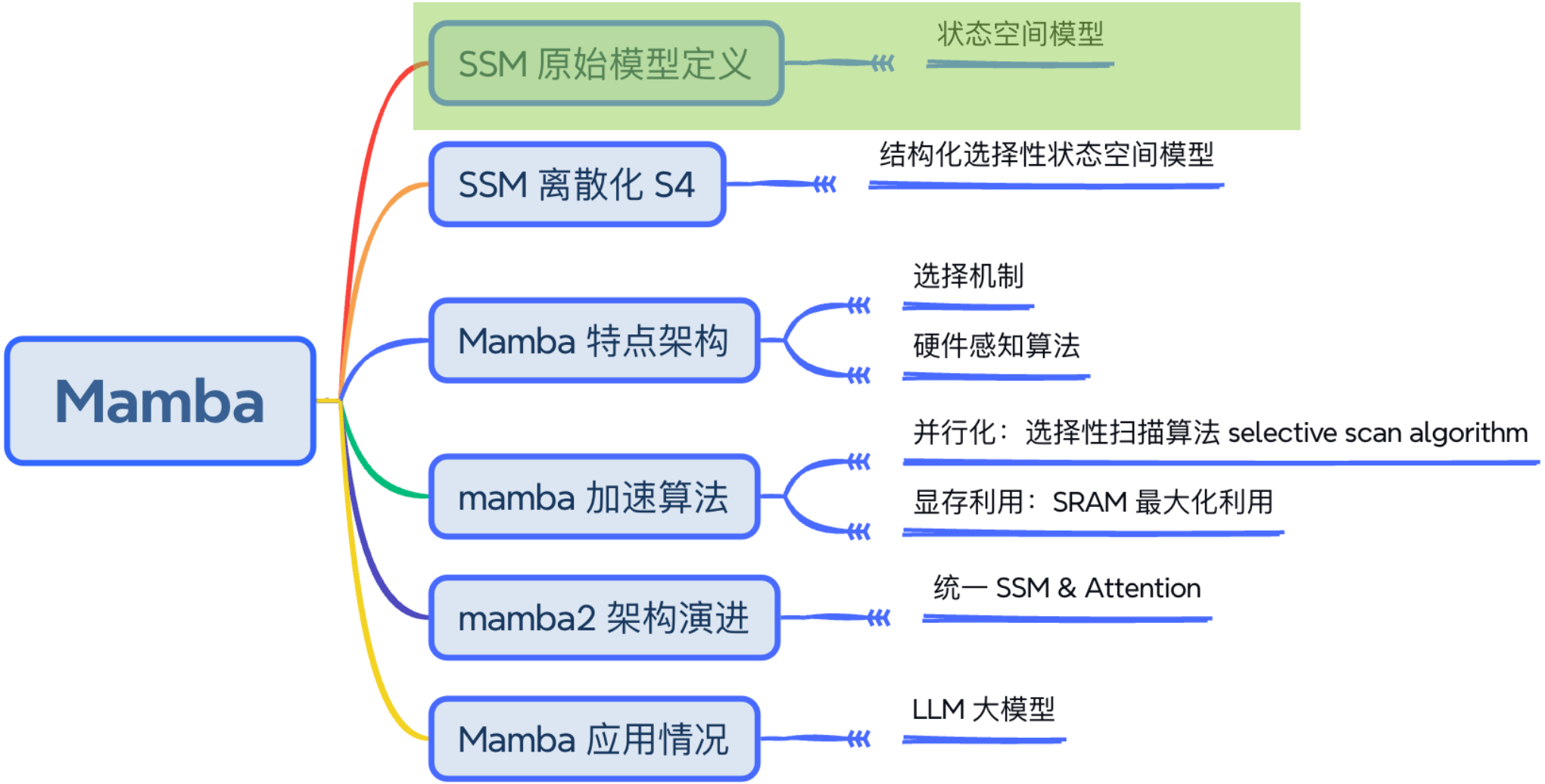
1. https://arxiv.org/abs/2006.16236
2. https://www.rwkv.com/
3. https://arxiv.org/abs/2105.14103
4. https://github.com/BlinkDL/RWKV-LM
5. https://wiki.rwkv.com/

02 Mamba

Mamba

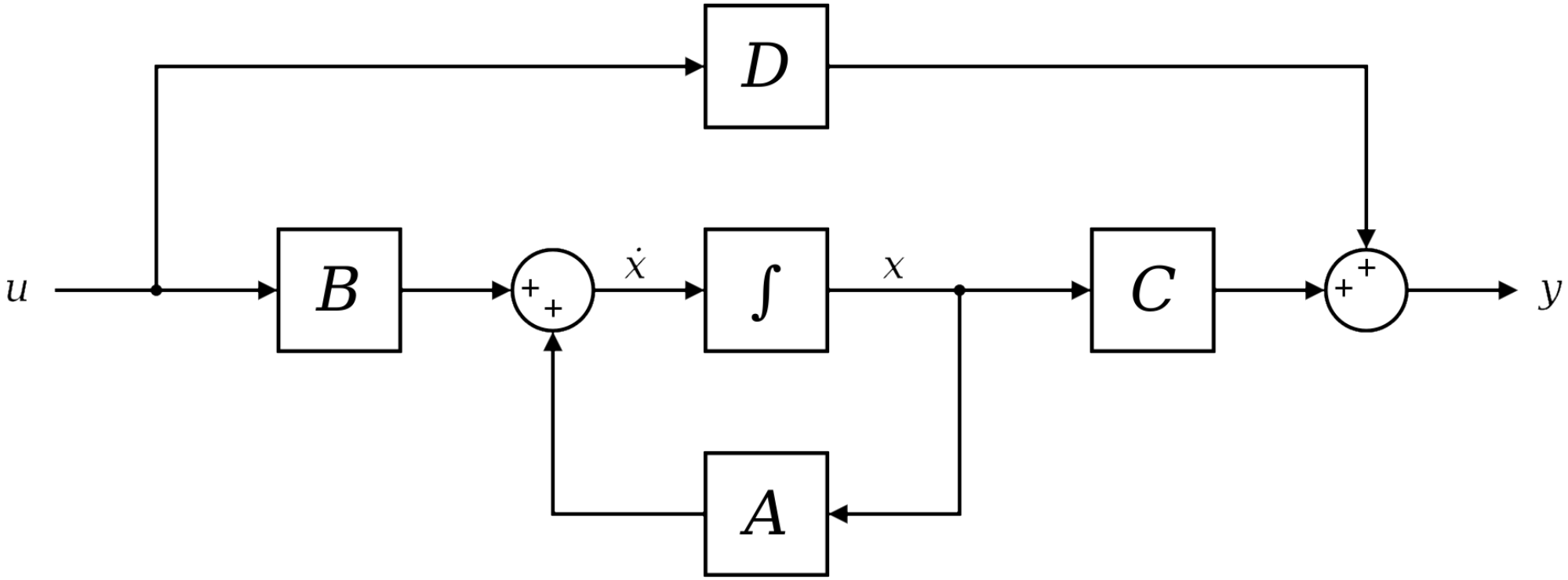

Mamba
SSM 原始模型定义
状态空间模型
SSM 离散化 S4
结构化选择性状态空间模型
Mamba 特点架构
选择机制
硬件感知算法
mamba 加速算法
并行化：选择性扫描算法 selective scan algorithm
显存利用：SRAM 最大化利用
mamba2 架构演进
统一 SSM & Attention
Mamba 应用情况
LLM 大模型

Mamba

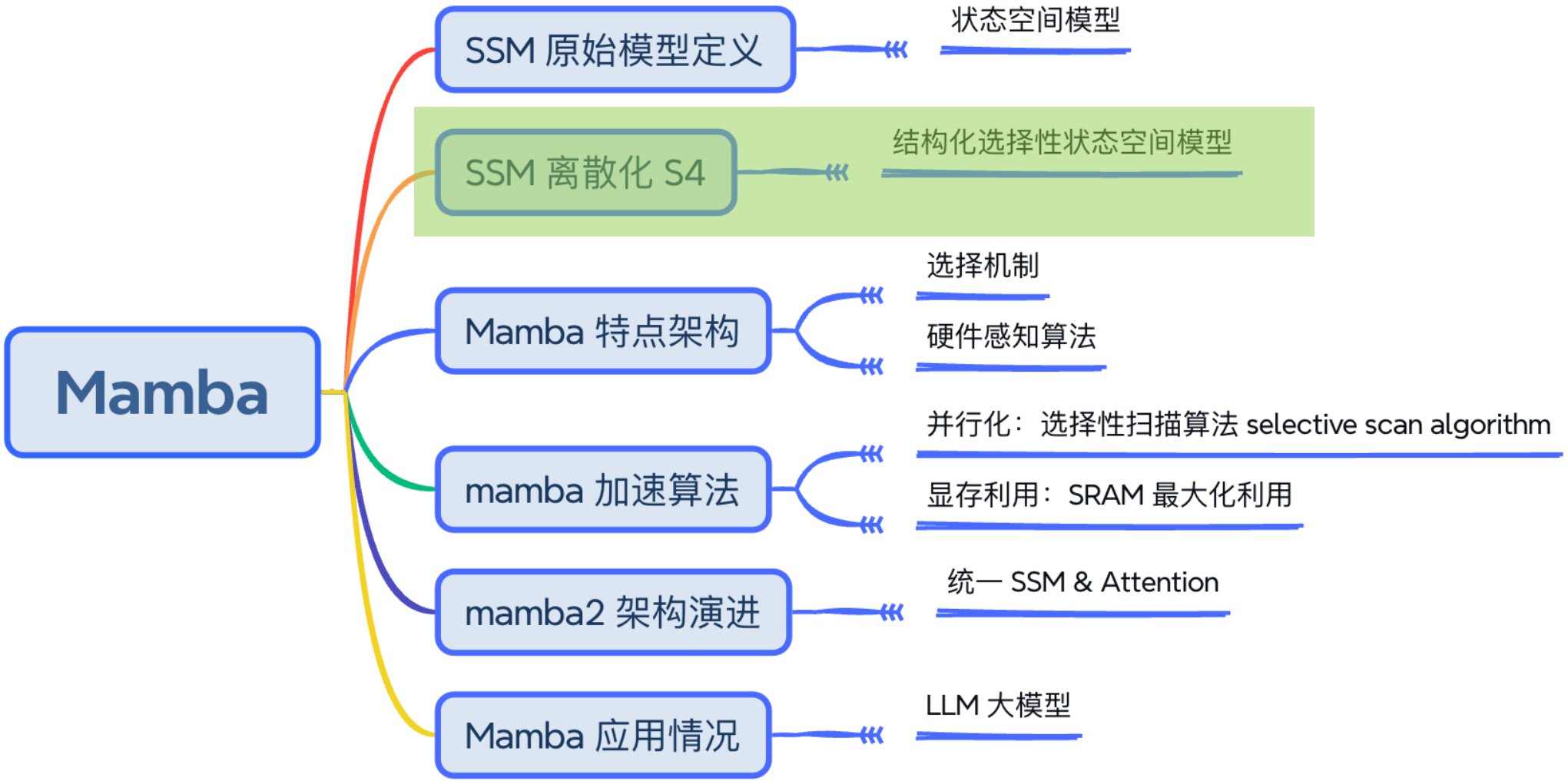

Mamba
SSM 原始模型定义
状态空间模型
SSM 离散化 S4
结构化选择性状态空间模型
Mamba 特点架构
选择机制
硬件感知算法
mamba 加速算法
并行化：选择性扫描算法 selective scan algorithm
显存利用：SRAM 最大化利用
mamba2 架构演进
统一 SSM & Attention
Mamba 应用情况
LLM 大模型

SSM

- Intelligent Control Systems: An Introduction with Examples
- State-space representation → State Space Model

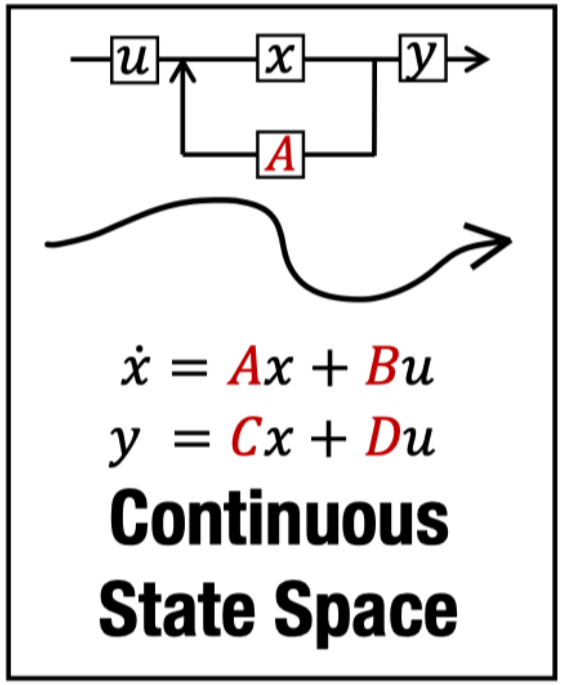


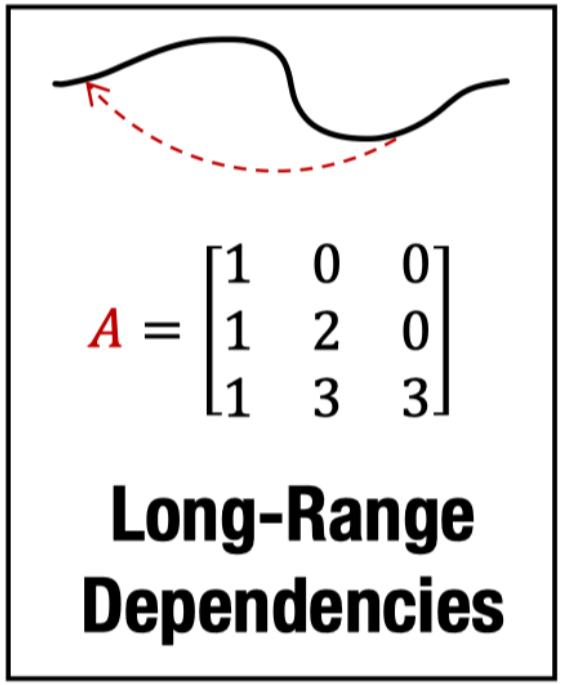
Mamba

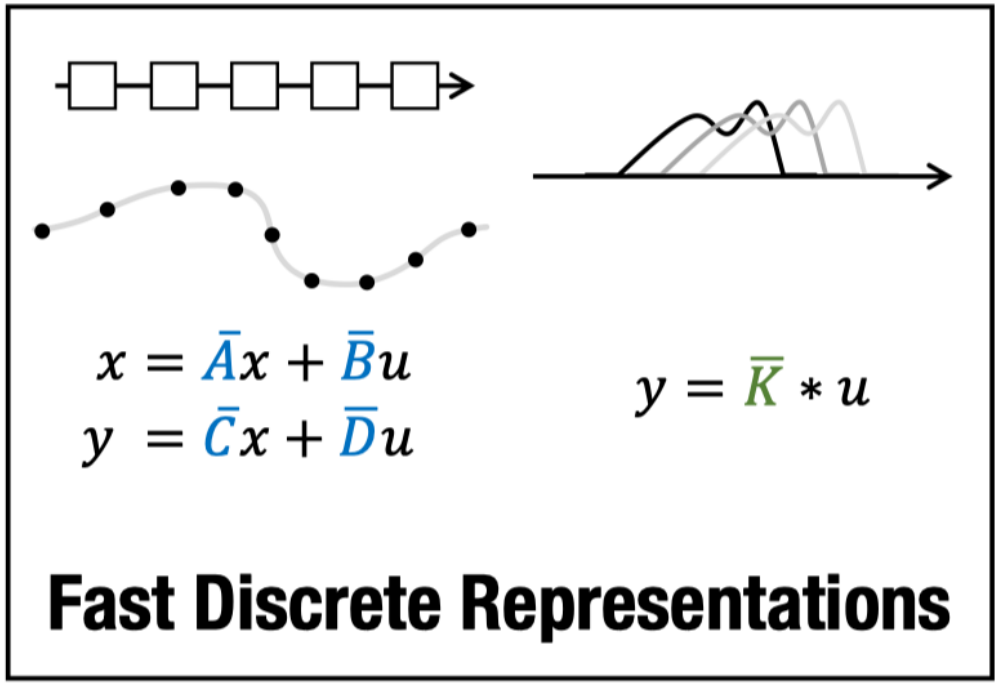

Mamba
SSM 原始模型定义
状态空间模型
SSM 离散化 S4
结构化选择性状态空间模型
Mamba 特点架构
选择机制
硬件感知算法
mamba 加速算法
并行化：选择性扫描算法 selective scan algorithm
显存利用：SRAM 最大化利用
mamba2 架构演进
统一 SSM & Attention
Mamba 应用情况
LLM 大模型

S4

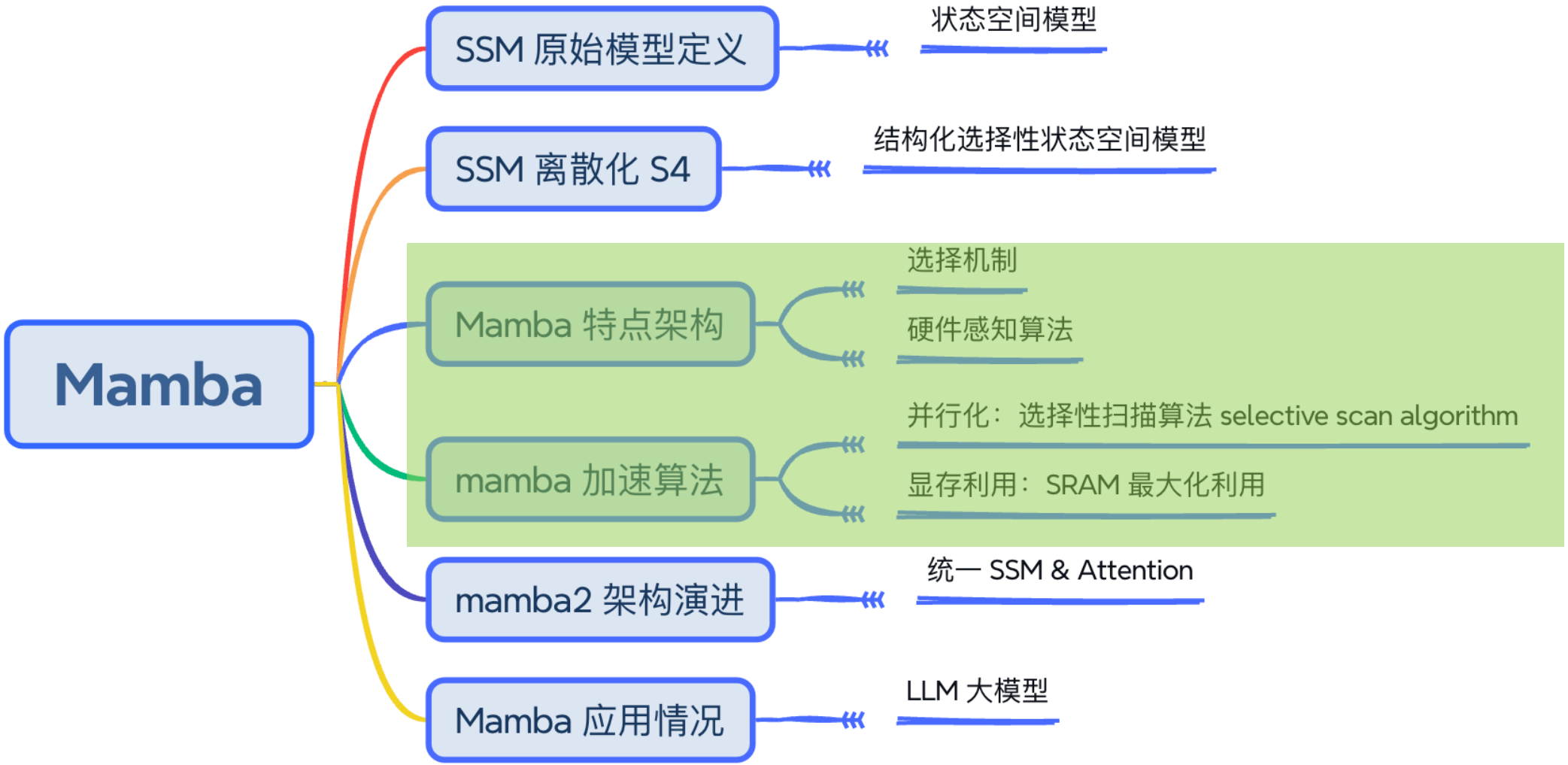
- Paper: Efficiently Modeling Long **S**equences with **S**tructured **S**tate **S**paces
- aka. Structured State Space sequence model (S4)







Mamba


Mamba
SSM 原始模型定义
状态空间模型
SSM 离散化 S4
结构化选择性状态空间模型
Mamba 特点架构
选择机制
硬件感知算法
mamba 加速算法
并行化：选择性扫描算法 selective scan algorithm
显存利用：SRAM 最大化利用
mamba2 架构演进
统一 SSM & Attention
Mamba 应用情况
LLM 大模型

mamba

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu*[1] and Tri Dao*[2]

[1]Machine Learning Department, Carnegie Mellon University

[2]Department of Computer Science, Princeton University
agu@cs.cmu.edu, tri@tridao.me

FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness

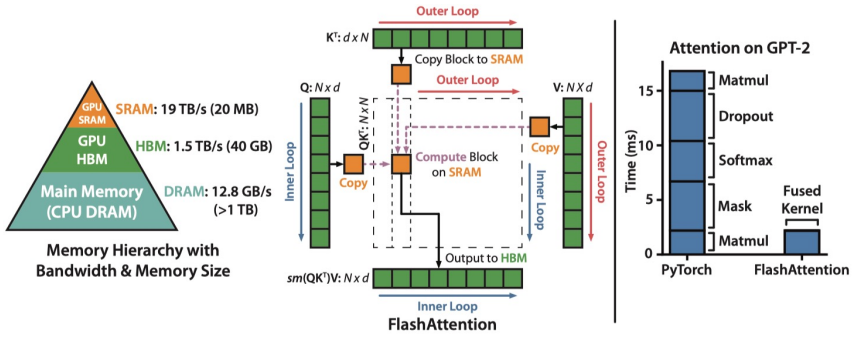
Tri Dao†, Daniel Y. Fu†, Stefano Ermon†, Atri Rudra‡, and Christopher Ré†

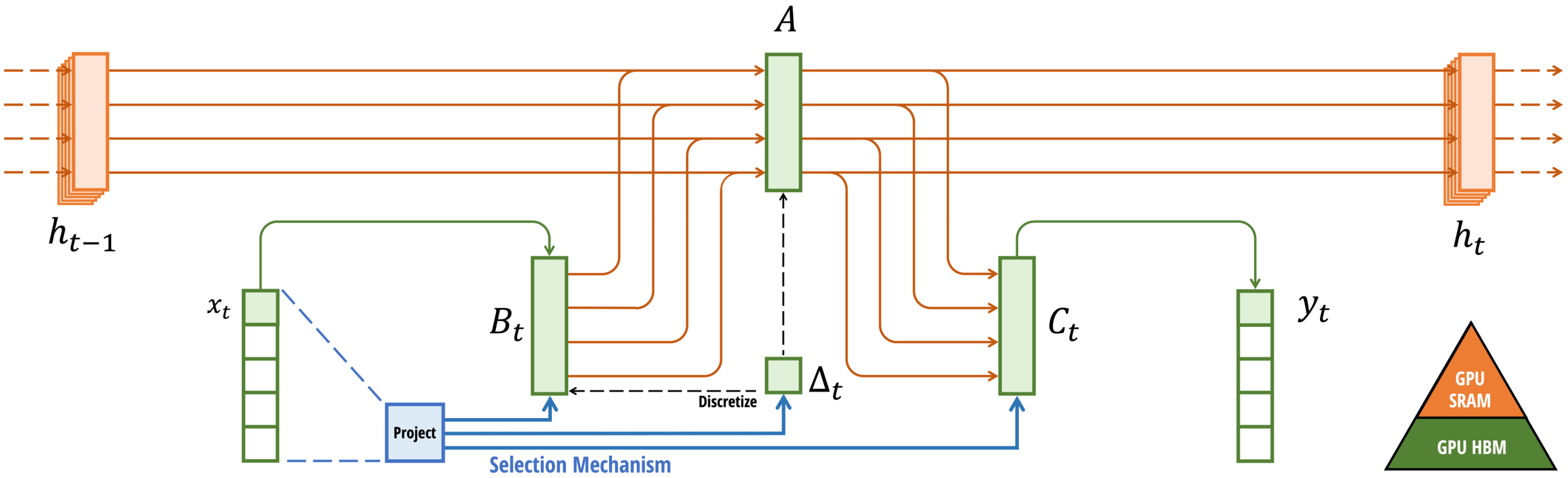
†Department of Computer Science, Stanford University
‡Department of Computer Science and Engineering, University at Buffalo, SUNY
{trid,danfu}@cs.stanford.edu, ermon@stanford.edu, atri@buffalo.edu,
chrismre@cs.stanford.edu

mamba

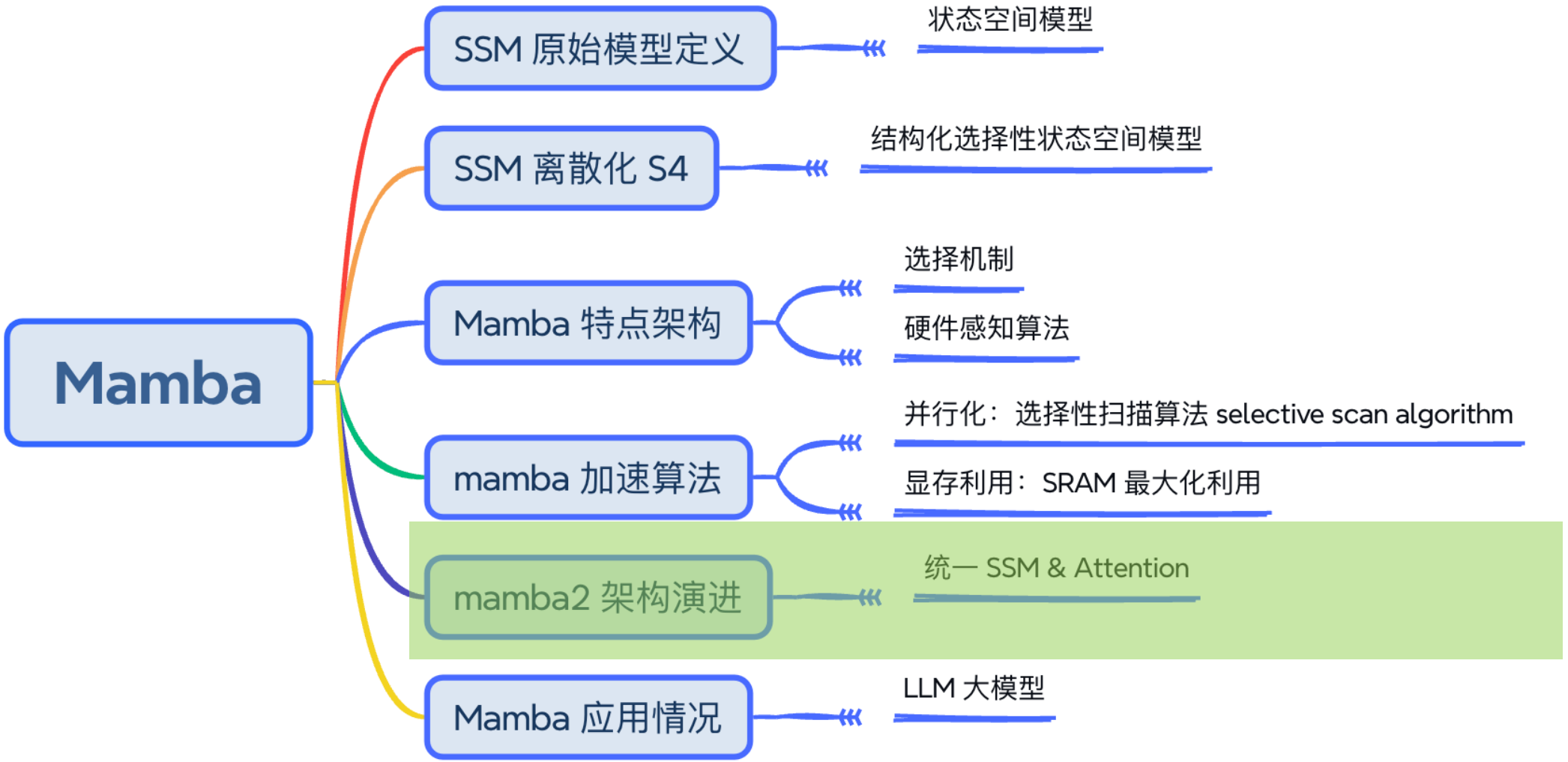
Selective State Space Model

with Hardware-aware State Expansion





Mamba


Mamba
SSM 原始模型定义
状态空间模型
SSM 离散化 S4
结构化选择性状态空间模型
Mamba 特点架构
选择机制
硬件感知算法
mamba 加速算法
并行化：选择性扫描算法 selective scan algorithm
显存利用：SRAM 最大化利用
mamba2 架构演进
统一 SSM & Attention
Mamba 应用情况
LLM 大模型

mamba2

Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality

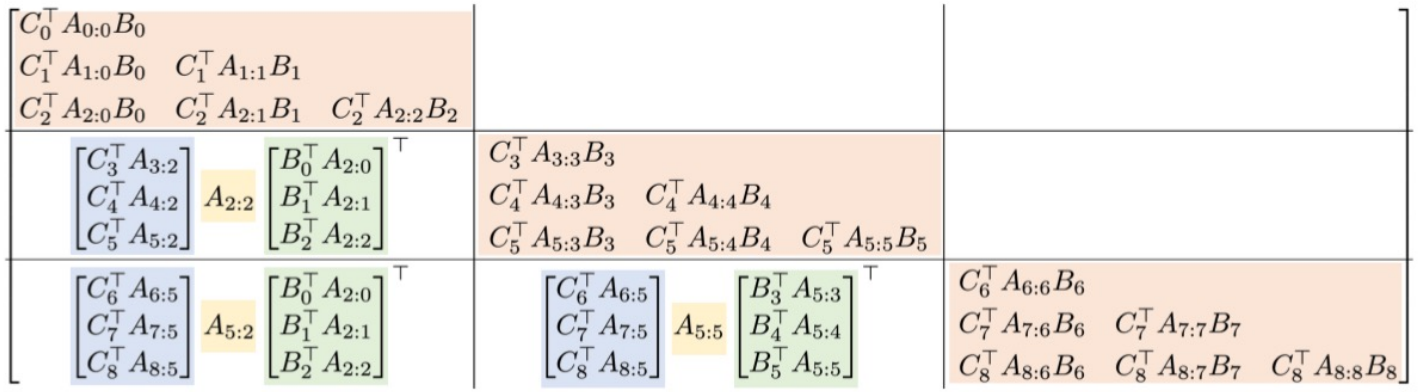
Tri Dao*[1] and Albert Gu*[2]

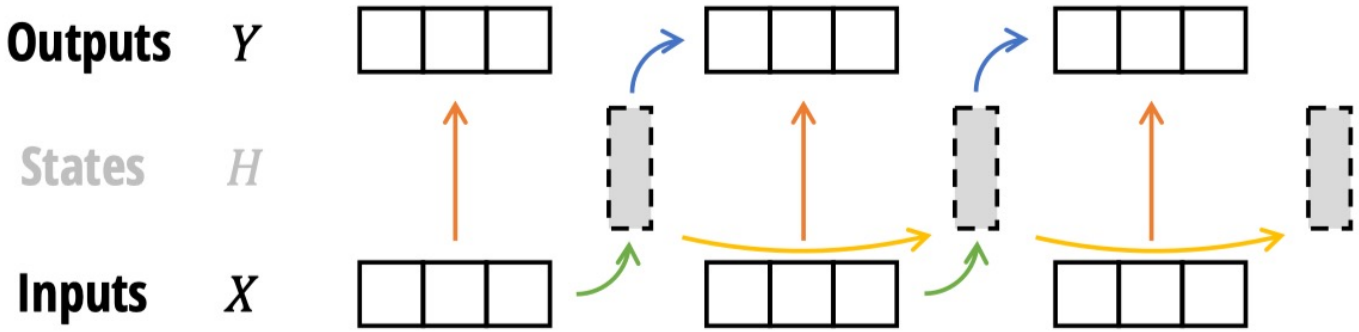
[1]Department of Computer Science, Princeton University

[2]Machine Learning Department, Carnegie Mellon University

tri@tridao.me, agu@cs.cmu.edu

mamba2





Semiseparable Matrix M

Block Decomposition

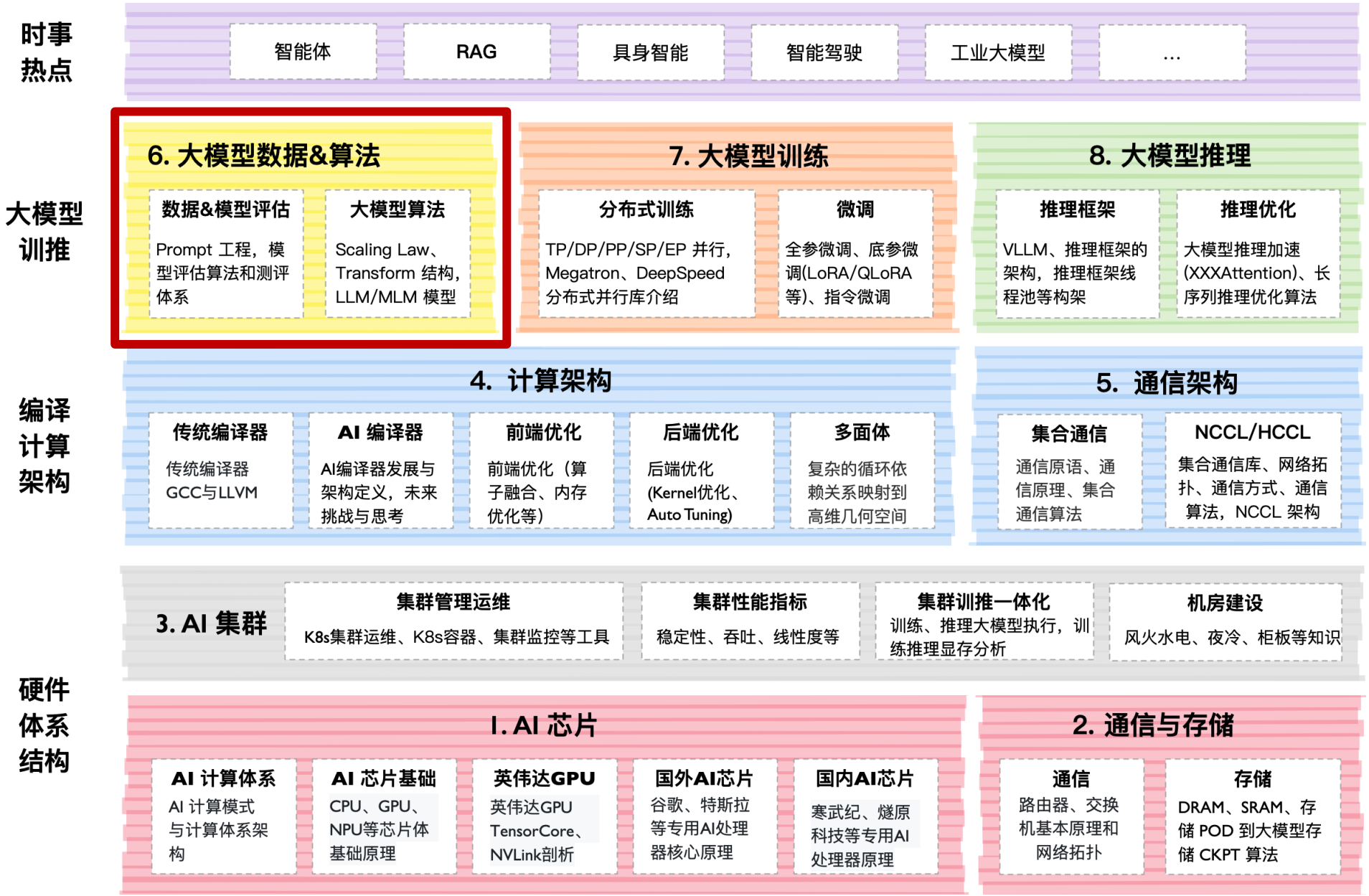
- Diagonal Block: Input → Output
- Low-Rank Block: Input → State
- Low-Rank Block: State → State
- Low-Rank Block: State → Output

Reference 参考&引用

1. https://srush.github.io/annotated-s4/
2. https://huggingface.co/blog/lbourdois/get-on-the-ssm-train
3. https://en.wikipedia.org/wiki/State-space_representation
4. https://arxiv.org/abs/2405.21060
5. https://github.com/state-spaces/mamba?tab=readme-ov-file

03

小结与思考


时事热点
智能体
RAG
具身智能
智能驾驶
工业大模型
...
大模型训推
6. 大模型数据&算法
数据&模型评估
Prompt 工程，模型评估算法和测评体系
大模型算法
Scaling Law、Transform 结构，LLM/MLM 模型
7. 大模型训练
分布式训练
TP/DP/PP/SP/EP 并行，Megatron、DeepSpeed 分布式并行库介绍
微调
全参微调、底参微调(LoRA/QLoRA 等)、指令微调
8. 大模型推理
推理框架
VLLM、推理框架的架构，推理框架线程池等构架
推理优化
大模型推理加速(XXXAttention)、长序列推理优化算法
编译计算架构
4. 计算架构
传统编译器
传统编译器 GCC与LLVM
AI 编译器
AI编译器发展与架构定义，未来挑战与思考
前端优化
前端优化（算子融合、内存优化等）
后端优化
后端优化(Kernel优化、Auto Tuning)
多面体
复杂的循环依赖关系映射到高维几何空间
5. 通信架构
集合通信
通信原语、通信原理、集合通信算法
NCCL/HCCL
集合通信库、网络拓扑、通信方式、通信算法，NCCL 架构
硬件体系结构
3. AI 集群
集群管理运维
K8s集群运维、K8s容器、集群监控等工具
集群性能指标
稳定性、吞吐、线性度等
集群训推一体化
训练、推理大模型执行，训练推理显存分析
机房建设
风火水电、夜冷、柜板等知识
1. AI 芯片
AI 计算体系
AI 计算模式与计算体系架构
AI 芯片基础
CPU、GPU、NPU等芯片体基础原理
英伟达GPU
英伟达GPU TensorCore、NVLink剖析
国外AI芯片
谷歌、特斯拉等专用AI处理器核心原理
国内AI芯片
寒武纪、燧原科技等专用AI处理器原理
2. 通信与存储
通信
路由器、交换机基本原理和网络拓扑
存储
DRAM、SRAM、存储 POD 到大模型存储 CKPT 算法

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem