An 88.36 TOPS/W

每瓦特执行 88.36 兆次 整数/浮点 计算

External Memory Access. EMA

(integer-float) Mixed-Precision

General Matrix Multiplication

Bit level weight compressed

weight 使用 bit 级压缩 — Active-Bit-Allocate-Format

Large Language Model Accelerator

LLM 加速器.

专门处理自回归模型 GPT. LLaMA 的 高记忆计算需求

with cluster Alligned INT-FP-GEMM.

整数,浮点数混合计算 矩阵乘法 — CAMP-PE. exponent 接近的 activation 分簇 转为整数计算

and Bi-Dimensional Workflow Reformulation

双维度重构 线性.非线性混合 — Bi-dimensional MAC-SFU Reformulation.

非线性计算 —taylor展开→ MAC. Op.

auto regressive: 生成与之前的生成有关

transformer → self Attention. 理解上下文

inference: auto regressive 方式 g. token. 基于已生成记录 生成新的 token

KV cache 的作用:

"The cat" 输入. —Encode→ model → token1 → "is" —计算 K.V 加入→ KV cache.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V$$

The cat

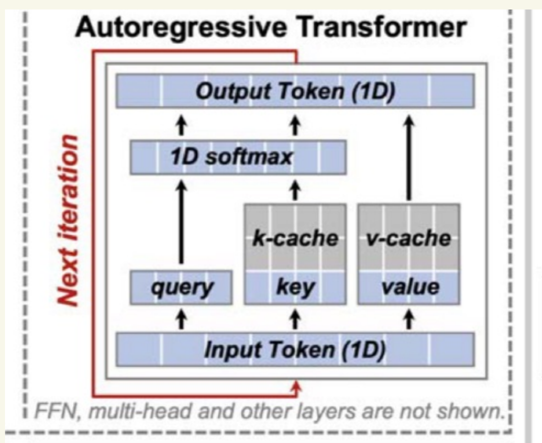
Q: 当前输入 Token 的 Query 向量

kv cache { K: 历史 token的 key 向量; V: 历史 token的 value向量 }

加入 输入

"The cat" "is" Input

KV cache



is 生成 Q. + KV cache

↓

attention 计算

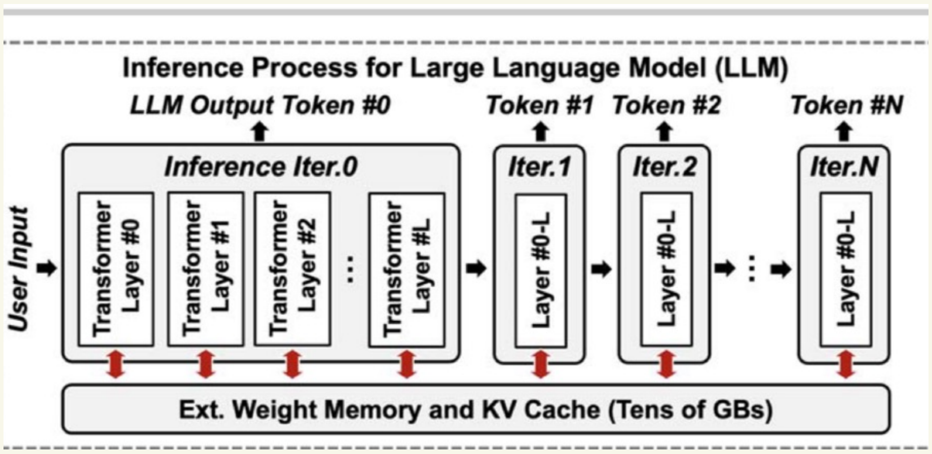
对于多层 Transformer. kv cache 会存储每层的 KV

inference: current_token → Query

all layers' k.v → K. V

attention compute.

new token store into kv cache.



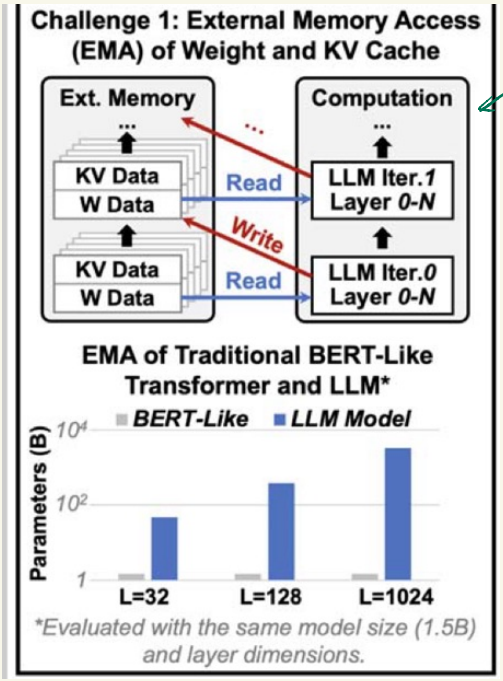
Mixed precision - General Matrix Multiplication

accuracy → weight → INT ?
activation → FP

weight → FP , → Energy Consumption 4.6x.

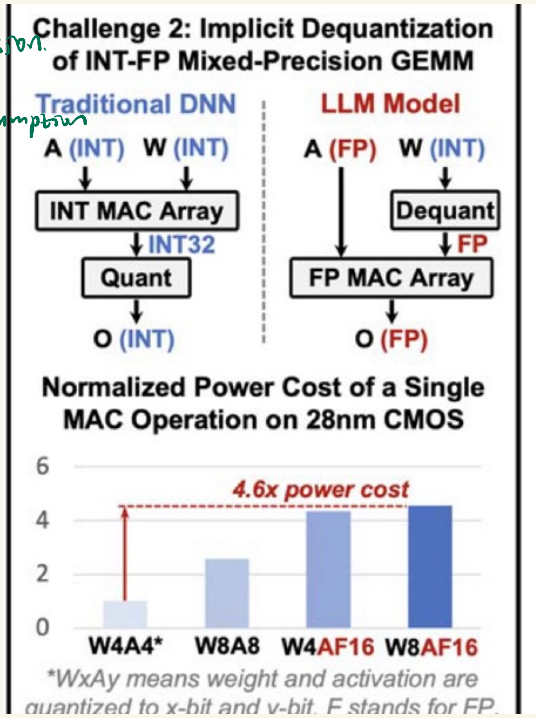
Special Function Unit.
非线性函数计算慢 , GEMM等待 SFU计算. 占用时间长

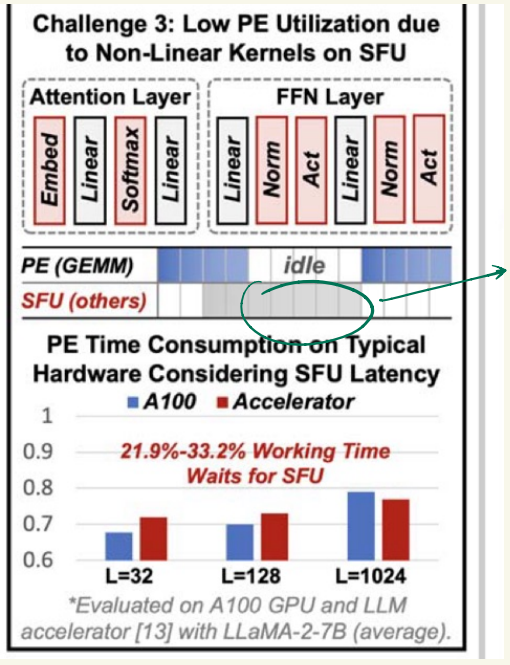
→ EMA过高 . 混合精度难实现. 非线性运算限制



auto regressive
kv. cache.
↓
EMA

mix precision.
energy consumption
complex





high latency

→
Accelerator: ① Active-Bit Allocate Format ABAF

reduce EMA for unimportant KV cache / weight Fetch / 91% weight is unimp.

approach: restrict "1" bits less than 3, + use low-width combination number.

codec-free.

reduce parameter size.

带宽压力↓.

② Cluster-Aligned Mixed-precision-Processing Element

CAMP-PE 混合精度聚类处理

split operands into clusters by exponent.

turn into INT GEMM. low energy consumption

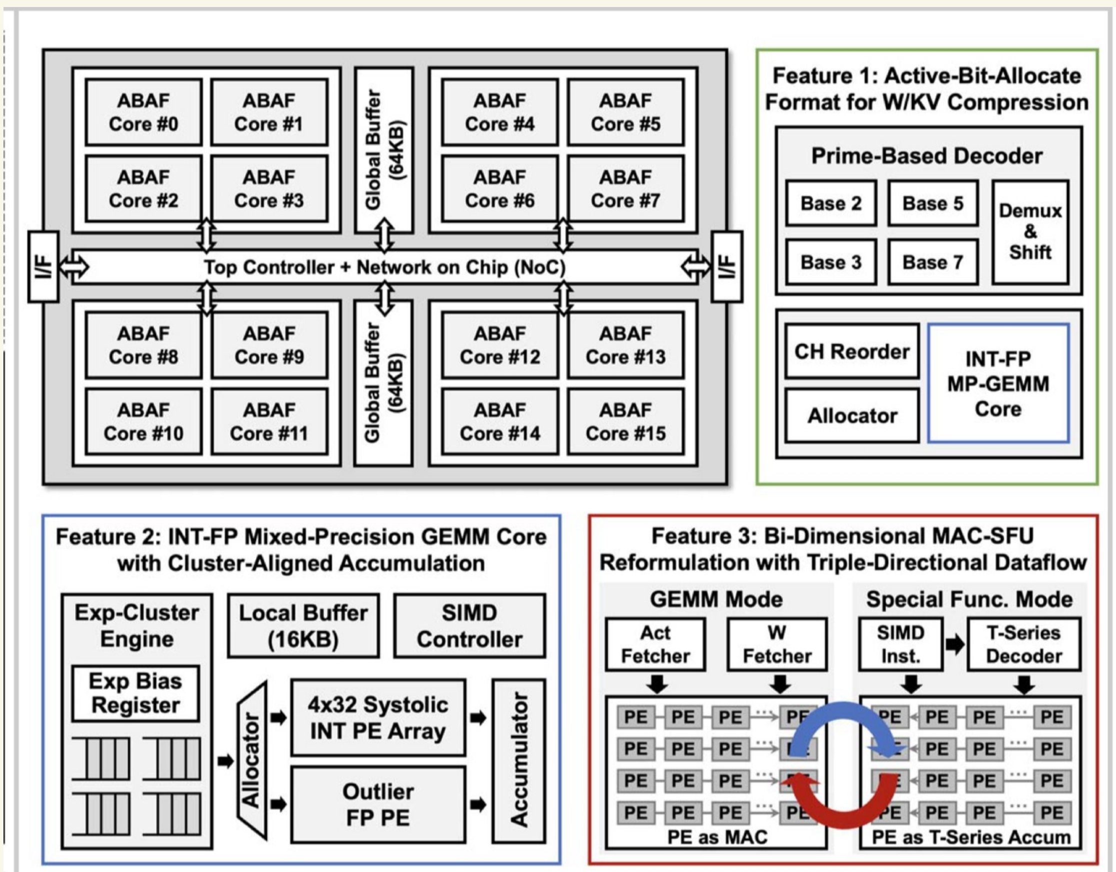
③ Bi-dimensional MAC-SFU Reformulation.

horner + taylor → (nonlinear) → MAC Op.

PE wait ↓

utilization↑

data transportation between linear and unlinear Units



network on chip.

用于芯片内部通信的结构化网络架构

取代了 bus 或点对点连接.

广泛用于 多核处理器 AI加速器

SoC, FPGA

NoC:

Node: 每个核/存储/处理单元接入 NoC 为 Node.

Route/switch: 连接 Node 两个, 决定 packet 路径

Link: Node 之间的链路.

packet: 数据划分为小块传输

拓扑结构: Mesh. # Torus

Ring. Tree. Crossbar

architecture of all chip:

ABAF ×16.

Global Buffer 128KB.

Network on chip + top Controller.

ABAF 包含16个 tile.

①使用 ABAF压缩权重与KV. 通过基于质数的电路实现无需查表解码的解压缩.

prime based decoder

CH recoder + allocator —output→ MP-GEMM

② feature2: exponent cluster 指数对齐引擎 将有相似指数的 activations 分组

1个CAMP-PE
- Local Buffer
- SIMD controller
- 4×32 INT PE.
- Outlier accumulator 异常值累加器

act cs

采用 bit serial MAC 逐位串行乘加.

③ BMSR 调度整个核心. 将非线性计算→GEMM.

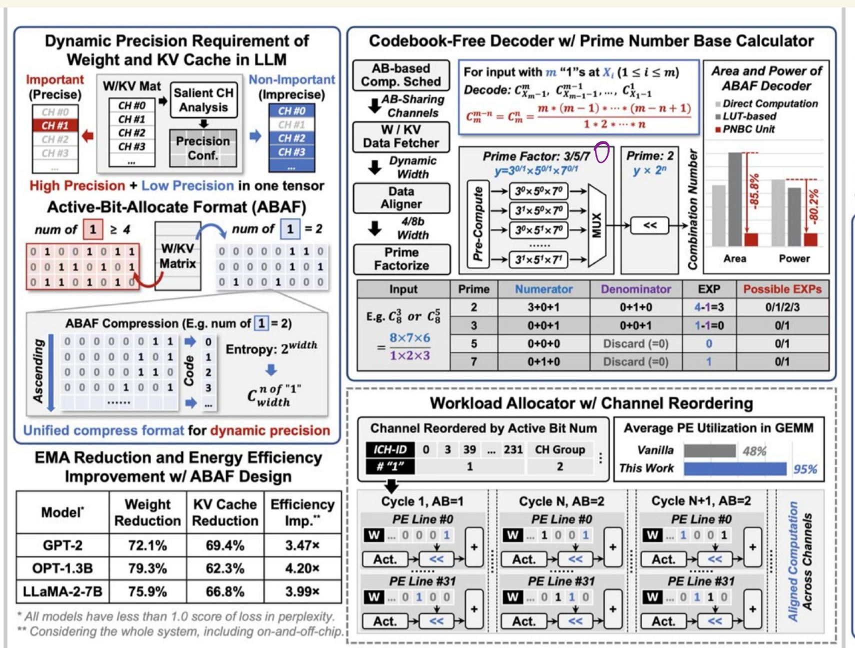


Figure 23.8.3: Principle of the proposed active-bit-allocate format (ABAF) and design details of codebook-free decoder and bit-serial workload allocator.

ABAF具体设计: 为1的称为 active bits.

在不重要的通道中, 只能有1-3个 ABs. 重要通道可以更多 ABs.

降低了信息熵

$C_m^n = C_{m-1}^n + C_{m-1}^{n-1}$

使用 Prime Number Based Calculator 解码.

不使用 codebook 的传统方法

对于 8bits. 只有包含质数 2, 3, 5, 7

2可以3次. 其余最高1次

用 multiplexer 选出结果 + shifter 移位计算 最终结果

用于快速还原原始组合.

Work flow:

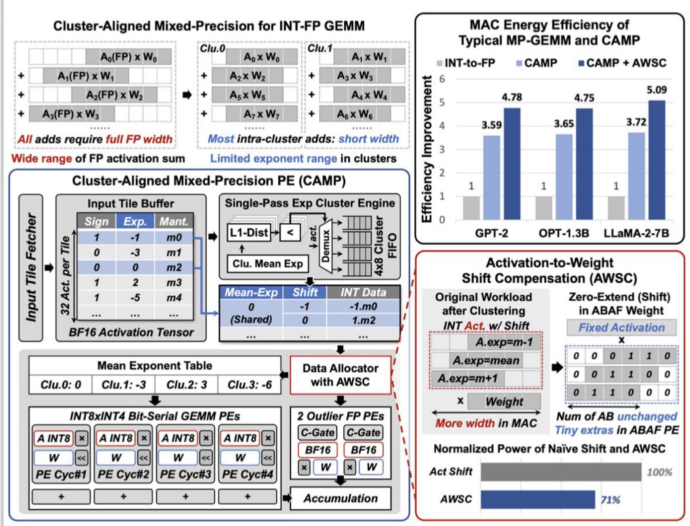


Figure 23.8.4: Cluster-aligned mixed-precision PE (CAMP) design for efficient integer-float MP- GEMM implementation.

CAMP-PE 设计:

将 activation 和权重的 ABAF 拉建成 KV组 根据 L1 距离分为4组

分别 储存在 4个深度为8的 FIFO 中. 每组内的指数差异小于3

转为整数操作.

指数对齐? → MAC 宽度增加.

activation-to-weight shift compensation.

activation → zero-extension.

没有增加计算负担. 因为 ABs 不变.

L1 distance: $\Sigma|x_i - y_i|$

workflow:

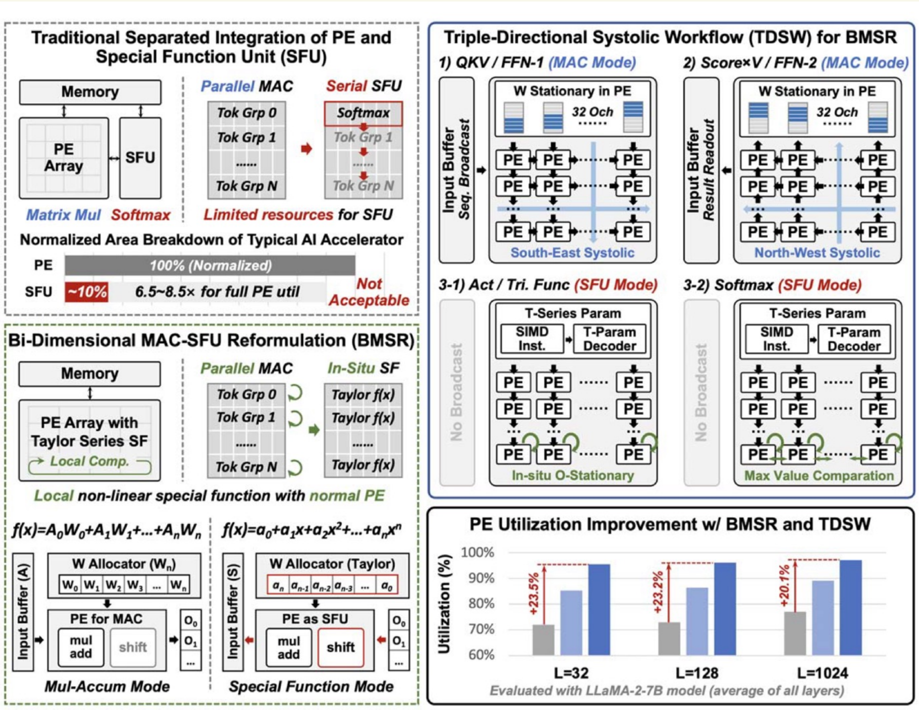


Figure 23.8.5: Bi-dimensional MAC-SFU reformulation design and triple-directional systolic workflow for high-utilization nonlinear function.

BMSR 设计:

taylor 展开 的计算形式与 GEMM 基本一致.

只需多个 2选1

可以 非线性计算 由 PE执行 不用 SFU

LLM有多种计算核. → 数据移动多

Triple - Directional - Systolic - workflow

数据传输方向 { in-situ. accumulation. / South east / north west

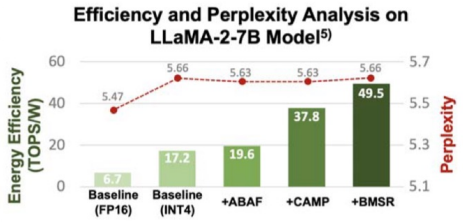
PE间插入比较器

\+

TDSW: 减少数据搬运开销

workflow:

Model	GPT-2-Small	OPT-1.3B	LLaMA-2-7B
Dataset	Wikitext-2 Language Modeling		
Accuracy and Loss[1)]	30.85 (+0.90)	15.50 (+0.88)	5.66 (+0.19)
Precision[2)]	A: BF16, W:INT8 with ABAF		
Latency (ms)	40.3 (6.72)[3)]	419.4 (68.9)	2139 (356.5)
Efficiency (uJ/Token)[4)]	3.11	36.2	194.1
Energy Saving	4.00×	4.62×	4.32×



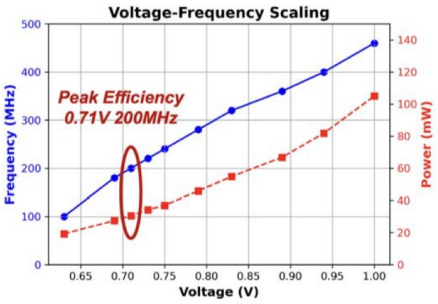
1) Perplexity on Wikitext-2 dataset, lower is better. Accuracy baseline is based on FP16 weight precision. 2) ABAF is used for weight dynamic compression. The best compression results of weight within 1.0 score of perplexity degradation limitation are used. 3) Evaluated with 6 chips working in parallel for actual working scenario, which has an total area of 21.12 mm², similar to SOTA LLM processor [23,24]. 4) System level evaluation with EMA included. 5) Evaluated on 1.0V, 460MHz, averaged on the generation process of 1024 tokens. Considering only on-chip performance and energy cost.

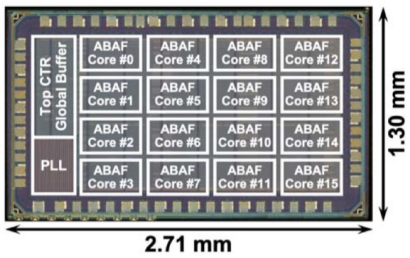
	ISSCC'22 [13]	VLSI'22 [20]	ISSCC'23 [22]	ISSCC'23 [17]	ISSCC'24 [23]	This Work
LLM Acceleration	NO	NO	NO	NO	Specific Model[5)]	General LLM Mem + Comp
Technique (nm)	28	5	12	28	28	28
Die Area (mm²)	6.82	0.153	4.60	3.93	20.25	3.52
Supply Voltage (V)	0.56-1.1	0.46-1.05	0.62-1.0	0.64-1.03	0.7-1.1	0.63-1.0
Frequency (MHz)	50-510	152-1760	77-717	20-320	50-200	50-460
Precision	INT12	INT4/INT8	FP4/FP8	INT8	INT8/16	BF16 (A) INT8 (W)
Power (mW)	12.1-272.8	NA	10-122 (FP8)	8.27-250.65	47.5-469.2	16.95-105.0
Performance[1)] (TOPS)	0.52-4.07[2)]	1.8 (INT8)	0.734 (FP4) 0.367 (FP8)	0.49-33.3	3.41 (INT8)	0.70[7)]-3.58[8)]
Energy Efficiency[1)] (TOPS/W)	1.91-27.56[2)]	39.1 (INT8)[3)]	18.1 (FP4) 8.24 (FP8)[3)]	1.96-53.83[4)]	22.9-47.8 (INT8)[6)]	10.07[7)]-88.36[8)]

1) External memory access is excluded for fair comparison, 1 MAC = 2OPs. 2) Achieved when computing Score×V with 90% weak related tokens. 3) With 50% non-zero input density. 4) 92% sparsity in attention span. 5) This work relies on big-little network algorithm, which is not supported by the majority of LLM like LLaMA. 6) With 10% sampling ratio for zero-skipping. 7) ABAF, CAMP, and BMSR disabled, evaluated on one layer of LLAMA-2-7B. 8) The peak throughput / efficiency for a single layer (including attention and FFN) with ABAF, CAMP, and BMSR, ABAF compression ratio = 75%.

Figure 23.8.6: Measurement results and comparison with state-of-the-art processors.

LLM generated response examples with proposed methods on OPT-1.3B	
Input	**Response**
What is math?	*Math is the study of numbers and their relationships.*
What is artificial intelligence?	*Artificial Intelligence, basically, it is the ability of a computer program to think.*
Where is New York?	*New York, NY is located in the United States.*





Specification		
Technique (nm)		28
Die Area (mm²)		3.52
Supply Voltage (V)		0.63 - 1.0
SRAM (KB)		384
Frequency (MHz)		50 - 460
Data Precision		A: BF16 W: INT8[1)]
Power (mW)		16.95 - 105
Performance (TOPS)[2)]		0.70[3)] - 3.58[4)]
Area Efficiency (TOPS/mm²)[2)]		0.20[3)] - 1.02[4)]
Energy Efficiency (TOPS/W)[5)]	0.71V 200MHz	10.07[3)] – 88.36[4)]
	1.0V 460MHz	6.73[3)] – 59.03[4)]

1) With ABAF compression. 2) Evaluated at 1.0V, 460MHz, 1MAC = 2OPs. 3) Baseline with no optimizations. 4) The peak throughput / efficiency for a single layer (including attention and FFN) with ABAF, CAMP, and BMSR, ABAF compression ratio = 75%. 5) EMA energy cost is not included.

Figure 23.8.7: Large language model generation examples and chip summary.