국내외 인공지능 반도체에 대한 연구 동향

for 암호연구회

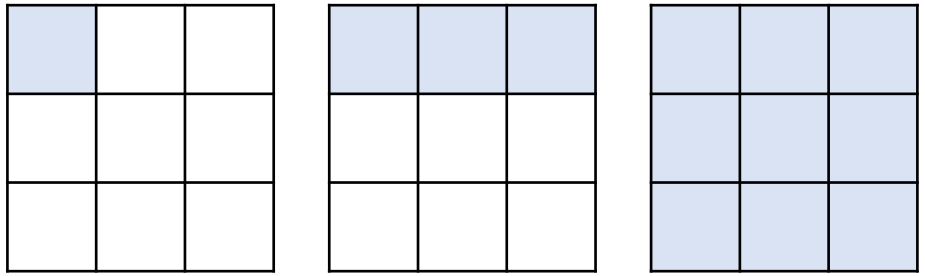
https://youtu.be/aarjphVOmY4

한성대학교 HANSUNG UNIVERSITY

CryptoCraft LAB

인공지능 반도체

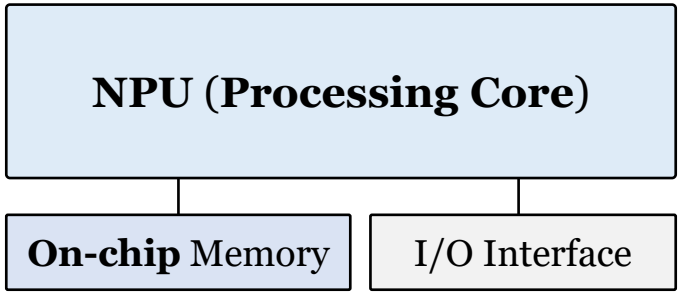
- CPU, GPU, NPU (인공지능 반도체)의 차이



- 크게 4종류가 있음
  - GPU (Graphics Processing Unit) : 그래픽 계산 위해 설계, 벡터 단위 연산을 병렬적으로 수행 가능
  - Tensor Cores : NVIDIA에서 개발, 행렬 곱 연산 가속화 (인공지능의 핵심 연산)
  - TPU (Tensor Processing Unit) : Google에서 개발한 텐서 연산에 최적화된 칩, 특히 TF와 함께 사용 시 효율적
  - NPU (Neural Processing Unit) : 인공지능에 필요한 연산에 특화된 칩, 위의 3개의 의미를 내포한다고 봐도 될듯
    전체 시스템의 에너지 효율성을 높이는데에 중점 (뒤에서 설명)
    딥러닝 연산엔 효율적이나, 다른 용도로 사용하기 어려움

인공지능 반도체

- 일반적인 구조
  - 연산을 위한 핵심 코어 존재
  - 온 칩 메모리 :연산 대상 데이터들을 올려 두고 빠르게 접근하면서 행렬곱 연산 수행
  - I/O Interface : 외부 데이터를 가져오는 통로가 됨
    - 해당 부분에서 많은 통신 오버헤드 발생하며, 이는 실제 딥러닝 연산의 100배 이상에 가까움



인공지능 반도체의 주요 특징

- 연산 단위
  - 각 인공지능 반도체는 연산 가능한 단위 및 특화된 연산 단위가 다를 수 있음
  - 딥러닝에 필요한 연산 단위를 제공함으로써 효율적 수행 가능
  - 연산 단위에 따라 대량의 데이터에 대한 연산을 수행하므로 인공지능 반도체의 성능 및 효율성을 결정하는 중요 요소

- 예전 세미나에서 다룬 내용 중에 양자화 라는 개념이 있음
  - 높은 정밀도의 데이터를 INT8 등의 정수 형태로 줄여버림으로써 연산을 빠르게 처리
  - 이는 추론을 위해 주로 사용됨
  - 이처럼 추론에 중점을 두고 INT만을 제공하는 칩도 존재

인공지능 반도체의 주요 특징

- 처리 속도
  - 딥러닝 모델의 학습 및 추론 속도에 직접적 영향을 미침

    →뒤에서 언급될 TOPS, TFLOPS 등의 인공지능 반도체의 성능 평가 지표에 영향을 줌

  - 빠른 처리 속도

    → 딥러닝 애플리케이션의 반응 속도 향상, 실시간 추론 가능, 대규모 데이터에 대한 연산 가속화

  - 특히! 대규모 작업이나 실시간 연산이 필요한 작업에서 중요

인공지능 반도체의 주요 특징

- 에너지 효율성
  - 굉장히 중요한 요소
  - 인공지능 반도체가 소비하는 전력 대비 수행 가능한 연산의 양
  - 딥러닝 작업은 대량의 데이터 처리 및 복잡한 계산이 필요 → 고성능을 유지 + 전력 소비를 최소화 필요
  - 주요 전력 소비 요인 : 딥러닝 연산 자체 + 외부 메모리에 접근하여 필요한 데이터를 칩에 올리는 과정
    - 대학원 수업에서 배웠던 부분인데, 외부 메모리에 접근하는 게 실제 연산보다 약 100배 이상의 전력이 소모된다고 함
      → 따라서 외부 메모리에 접근하는 횟수를 줄일 필요가 있음
  - 따라서 저전력 설계 및 에너지 효율적 연산 알고리즘들이 연구되는 중
  - 특히 에너지 효율적인 연산 알고리즘은 인공지능 반도체의 효율성 개선을 위해 필수적
    - 연산량을 줄이면서 정확도를 유지 또는 개선할 수 있어야 함
    - 이를 위해 딥러닝 아키텍처에 대한 연구들도 함께 진행 중
      → 이전 세미나에서 했던 내용들인데, 프루닝, 양자화 등의 방법을 통해 연산할 대상 자체를 간소화 하는 방법이 있음

인공지능 반도체의 주요 특징

- 결론
  - 가장 중요한 3요소 : 연산 단위, 처리 속도, 에너지 효율성
  - 해당 요소들은 딥러닝 시스템의 성능 및 운영 비용에 직접적 영향 미침
  - 따라서 특정 작업에 맞는 인공지능 반도체 선택 시 매우 큰 영향을 미침
  - 물론, 해당 3요소 외에도 메모리 용량, 설계 유연성 등의 요소들도 고려해야함

국내외 인공지능 반도체 개발 동향

- NVIDIA H100
  - 해당 칩에 탑재된 Tensor Core로 인해 딥러닝 학습 및 추론, 고성능 컴퓨팅, 컴퓨터 비전 등에 매우 적합
  - NVLink 스위치 시스템을 사용하면 최대 256개의 H100을 연결하여 작업을 가속화 가능
  - 특히, 딥러닝 작업에 주로 사용되던 NVIDIA V100에 비해 연산 속도가 5배 이상 향상

    → 매개변수가 조 단위인 언어모델을 처리 가능
  - 여러 데이터 타입에 대한 연산을 지원하고, 모든 데이터 타입에 대해 매우 높은 성능
  - 고성능의 실시간 추론 능력을 갖추었으나 이전 세대의 GPU보다 더 많은 전력 (W)을 소모한다는 단점 존재

국내외 인공지능 반도체 개발 동향

- META MTIA
  - 컨텐츠 추천을 위한 인공지능 모델을 적극적으로 활용 중이며, 이를 위한 칩을 설계 → MTIA
  - 대규모 트레이닝 및 추론을 위한 인공지능 가속기 (Meta Training and Inference Accelerator)

    → 칩 하나로 학습과 추론을 모두 처리하게 했다는 점에서 주목받고 있음
  - Pytorch 구현만 지원하며, 해당 프레임워크에 완전히 최적화 되도록 설계됨
  - 온 칩 메모리가 128MB로 굉장히 크기때문에 자주 액세스하는 데이터 및 명령에 대해 더 낮은 대기시간 소요
  - 에너지 효율성이 좋은 칩으로 알려져 있음

국내외 인공지능 반도체 개발 동향

- Tesla D1
  - 자율 주행 자동차에 완전 자율 주행 기능 (Full Self Driving, FSD)을 탑재하기 위해 D1을 개발
  - 요즘 주목 받고 있는 GPT 등과 같은 거대 딥러닝 모델보다는 자율 주행과 연관된 딥러닝 모델을 대상으로 함
    - 주로 비디오 데이터에 대한 학습 및 도로 위의 물체에 대한 자동 라벨링 등의 기능이 필요
  - BF16, FP8, INT8 등 다양한 데이터 타입을 처리할 수 있는 행렬 곱 전용 연산기와 벡터 연산기가 포함
  - 따라서, 고도로 병렬화 된 연산이 가능하며, 학습 및 추론 과정에 필요한 연산을 효율적으로 수행 가능하도록 설계됨

국내외 인공지능 반도체 개발 동향

- Apple Neural Engine
  - Apple이 자체 개발한 신경망 엔진
  - 기계 학습 모델을 빠르고 효율적으로 처리할 수 있도록 설계
  - 2017년 iPhone X에 탑재된 A11 칩의 일부로 처음 포함
    → A11은 최대 0.6 TFLOPS였으나 5세대 16-core ANE는 26배가량 높아진 15.8 TFLOP
    → 스마트폰에서도 높은 성능의 NPU를 사용할 수 있음을 보임
  - M2 pro chip에는 CPU 및 GPU와 함께 학습 및 추론 프로세스를 가속화하기 위한 16-core Neural Engine이 탑재
    → M1에도 탑재되어있으나, M1 대비 연산 처리 속도가 40퍼센트 향상되어 초당 최대 15.8 TFLOPS의 연산을 처리
  - 그러나 다른 NPU들에 비해 성능이 좋지는 않고, 노트북에 CPU, GPU와 함께 기본적으로 탑재되었다는 정도인 듯..

국내외 인공지능 반도체 개발 동향

- ETRI AB9
  - FP16 지원하며, 추론 연산 가속화를 위한 칩
  - 이동통신, 자율주행, 지능형 로봇 및 드론 등에 활용하기 위해 개발
  - 신경망의 레이어에 필요한 대부분의 연산 수행 가능하도록 되어있으며, 추가적인 과정 필요 없이 사용자가 쉽게 프로그래밍 가능
  - 개발 목적 상, 전력 소비가 낮고 속도가 어느 정도 보장되어야함
    - FP16 기준으로 15W의 낮은 전력을 소모하면서 초당 40 TFLOPS라는 높은 연산 능력을 보여준다는 장점
  - 향후 10W 이내의 전력 소비량을 달성하여 자율주행 자동차 등에 안전하게 탑재되는 것을 목적으로 함

국내외 인공지능 반도체 개발 동향

- SAPEON X220/X330
  - 국내 SK텔레콤에서 개발한 인공지능 반도체
  - 제안서 쓸 때까지는 X220만 있었던 것으로 기억
  - 학습시키고 추론 모두 가능하며, 음성 인식, 이미지 분석, 자연어 처리 등 여러 작업 가속화 가능
  - X220은 INT8만 지원 → 딥러닝 추론을 중점으로 두는 칩이기 때문
  - 전력 소모를 최소화하여 운영비용을 절감할 수 있도록 하는 것에 중점으로 둠
  - 대표적 언어인 Pytorch, Tensorflow와 호환되게 함으로써 딥러닝 기술의 대중화 및 상용화를 촉진시키는 것을 목적으로 함

  - X220의 다음 모델인 X330은 부동 소수점 연산 (FP16, FP8)도 지원하고, 정수는 INT8만 지원
  - X220과 비교하였을 때 4배 이상의 연산 성능, 2배 이상의 전력 효율 달성
  - 2025년에는 X430 공개를 목적으로하며, 해당 칩은 초거대 딥러닝 모델을 타겟으로 함

국내외 인공지능 반도체 개발 동향

- FuriosaAI Warboy
  - 클라우드 컴퓨팅, 엣지 컴퓨팅 등과 같이 여러 환경에서의 인공지능 작업을 가속화 하는 데에 초점
  - 해당 칩은 EfficientNetV2 (CNN 기반의 효율적인 대표적인 딥러닝 모델)와 같이 경량화 모델에 초점을 둔 것으로 보임
    - NVIDIA의 T4와 비교하였을 때, 2.83배의 와트 당 처리량 달성
  - 높은 전력 소모를 최소화 하고, 양자화를 통해 처리속도 및 성능을 최적화

국내외 인공지능 반도체 개발 동향

- Rebellions ATOM
  - 최대 16개의 작업을 동시에 지원 가능
  - 효율적인 하드웨어 구현 → 소규모 어플리케이션부터 대규모 작업까지 효율적으로 가속화 가능하다고 함
  - 데이터 통신 오버헤드 등과 같은 문제점을 극복하기 위해 에너지 효율성을 위한 저전력 설계 방법론을 적용
  - CNN, LSTM, BERT, 최신 트랜스포머 모델 (T5, GPTs 등)을 포함한 다양한 유형의 신경망을 효율적으로 가속화 가능

국내외 인공지능 반도체에 대한 비교 분석

- **TOPS** : 초당 테라 연산 처리량
- **TFLOPS** : 초당 테라 부동소수점 연산 처리량

- TOPS가 더 넓은 범위지만, 주로 TFLOPS로 성능이 제공되고 있으며, TOPS는 정수 연산에 대한 성능 측정 시 주로 사용
  → **여기선 두 지표를 혼용**

- **연산량, 전력, 온 칩 메모리, 에너지 효율성에 초점을 두어 평가**

- 오른쪽 표에서 볼 수 있듯이, **연산량과 전력은 트레이드 오프**가 존재
  → **이 두 요소를 모두 고려한 에너지 효율성에 대한 평가가 필요**

- 또한, 향후 AI 칩 개발에 있어 에너지 효율성 향상을 위한 전력과 연산량의 균형이 중요할것으로 생각됨
  → 현재로서는 대부분 높은 연산량에 초점

표 3 국내외 인공지능 반도체에 대한 비교 분석 (연산 단위, 연산 속도, 전력, 온 칩 메모리)

제품명	제조사	Data Types	TOPS	TFLOPS	Watt(W)	On-Chip Memory
H100	NVIDIA (국외)	FP64, TF32, FP32, FP16, FP8, INT8	3958 (INT8)	34 (FP64), 37 (FP32), 989 (TF32), 1979 (FP16), 3958 (FP8)	>700	116MB
MTIA	Meta (국외)	FP16, BF16, INT8	102.4 (INT8)	51.2 (FP16)	25	128MB
D1	Tesla (국외)	FP32, BF16, FP8, INT8	-	22.6 (FP32), 362 (BF16)	400	442.5MB
Neural Engine	Apple (국외)	FP16	-	15.8 (FP16)	-	-
AB9	ETRI (국내)	FP16	-	40 (FP16)	15~40	>36MB
X220	SAPEON (국내)	INT16, INT8, INT4	87 (INT8, Compact), 174 (INT8, Enterprise)	-	65 (Compact), 135(Enterprise)	-
X330		FP16, FP8, INT8	-	184 (FP16, Compact), 367 (FP8, Compact), 368 (FP16, Prime), 734 (FP8, Prime)	75~120 (Compact), 250 (Prime)	-
Warboy	Furiosa AI (국내)	INT8	64 (INT8)	-	40~60	32MB
ATOM	Rebellions (국내)	FP16, INT8, INT4, INT2	128 (INT8)	32 (FP16)	60~150	64MB

국내외 인공지능 반도체에 대한 비교 분석

*연산 단위가 다르므로 완전히 공정한 비교는 불가
가장 많이 지원하는 연산 단위인 FP16과 INT8을 대상으로 비교

- **연산 단위**
  - **NVIDIA H100**이 **가장 많은 데이터 타입을 지원**

    → 가장 범용적 연산 가능할 것
  - 이에 비해 다른 인공지능 반도체들은 거의 **FP16, FP8, INT8을 지원**하며, FP를 지원하지 않는 경우도 있음
  - 이처럼 특정 작업/연산 단위에 특수화 된 칩들이 존재

- **연산 속도 (빠른 순서대로,** 전력 소모나 범용성 및 특수성은 고려하지 않음**)**
  - **H100 > X330, D1 > X220, ATOM, MTIA > AB9, Warboy**

- **전력 소모 (높은 순서대로)**
  - 에너지 효율성을 위해 중요한 요소, 개발 목적에 따라 전력에도 차이 존재
  - **H100 > D1 > X330 > ATOM > X220 > Warboy > AB9,MTIA**
  - AB9은 자율 주행을 위해 저전력 설계 된 칩 (빠른 연산 속도 지원 X)
  - D1도 자율 주행 타겟, 그러나 높은 전력 소모하지만 높은 연산 속도 가짐

    → 전력 소모는 높지만 더욱 안전한 자율 주행 가능함을 주장

표 3 국내외 인공지능 반도체에 대한 비교 분석 (연산 단위, 연산 속도, 전력, 온 칩 메모리)

제품명	제조사	Data Types	TOPS	TFLOPS	Watt(W)	On-Chip Memory
H100	NVIDIA (국외)	FP64, TF32, FP32, FP16, FP8, INT8	3958 (INT8)	34 (FP64), 37 (FP32), 989 (TF32), 1979 (FP16), 3958 (FP8)	>700	116MB
MTIA	Meta (국외)	FP16, BF16, INT8	102.4 (INT8)	51.2 (FP16)	25	128MB
D1	Tesla (국외)	FP32, BF16, FP8, INT8	-	22.6 (FP32), 362 (BF16)	400	442.5MB
Neural Engine	Apple (국외)	FP16	-	15.8 (FP16)	-	-
AB9	ETRI (국내)	FP16	-	40 (FP16)	15~40	>36MB
X220	SAPEON (국내)	INT16, INT8, INT4	87 (INT8, Compact), 174 (INT8, Enterprise)	-	65 (Compact), 135(Enterprise)	-
X330		FP16, FP8, INT8	-	184 (FP16, Compact), 367 (FP8, Compact), 368 (FP16, Prime), 734 (FP8, Prime)	75~120 (Compact), 250 (Prime)	-
Warboy	Furiosa AI (국내)	INT8	64 (INT8)	-	40~60	32MB
ATOM	Rebellions (국내)	FP16, INT8, INT4, INT2	128 (INT8)	32 (FP16)	60~150	64MB

국내외 인공지능 반도체에 대한 비교 분석

*연산 단위가 다르므로 완전히 공정한 비교는 불가
가장 많이 지원하는 연산 단위인 FP16과 INT8을 대상으로 비교

- **온 칩 메모리**
  - 외부 메모리 접근은 매우 큰 전력 소모를 야기하므로
    → 온칩 메모리가 크면 일반적으로 지연시간 감소, 연산 성능 증가
  - 온 칩 메모리의 크기를 늘리면: 추가 비용, 칩의 전력 소모 증가 등 발생
    따라서, 칩 설계 시 성능, 비용, 전력 소모 사이의 균형을 찾는 것이 중요
  - **D1 > MTIA > H100 > ATOM > AB9 > Warboy**

- **에너지 효율성**
  - 소비 전력 대비 수행 가능한 연산량 (**TOPS or TFLOPS per Watt**)
  - **FP16**
    - **H100 (2.82) > AB9 (2.66~1.00) 및 X330 (2.45~1.53) > MTIA (2.05)> D1 (0.91)>ATOM (0.53~0.21)**
    - H100이 높은 전력에도 불구하고 높은 에너지 효율성 달성
  - **INT8**
    - **H100 (5.65) > MTIA (4.10) > ATOM (2.13~0.85) 및 X220 (1.33, 1.28) > Warboy (1.06)**
  - 두 연산 단위에 대해 **H100이 에너지 효율성에서 가장 좋은 성능 달성**

표 3 국내외 인공지능 반도체에 대한 비교 분석 (연산 단위, 연산 속도, 전력, 온 칩 메모리)

제품명	제조사	Data Types	TOPS	TFLOPS	Watt(W)	On-Chip Memory
H100	NVIDIA (국외)	FP64, TF32, FP32, FP16, FP8, INT8	3958 (INT8)	34 (FP64), 37 (FP32), 989 (TF32), 1979 (FP16), 3958 (FP8)	>700	116MB
MTIA	Meta (국외)	FP16, BF16, INT8	102.4 (INT8)	51.2 (FP16)	25	128MB
D1	Tesla (국외)	FP32, BF16, FP8, INT8	–	22.6 (FP32), 362 (BF16)	400	442.5MB
Neural Engine	Apple (국외)	FP16	–	15.8 (FP16)	–	–
AB9	ETRI (국내)	FP16	–	40 (FP16)	15~40	>36MB
X220	SAPEON (국내)	INT16, INT8, INT4	87 (INT8, Compact), 174 (INT8, Enterprise)	–	65 (Compact), 135(Enterprise)	–
X330		FP16, FP8, INT8	–	184 (FP16, Compact), 367 (FP8, Compact), 368 (FP16, Prime), 734 (FP8, Prime)	75~120 (Compact), 250 (Prime)	–
Warboy	Furiosa AI (국내)	INT8	64 (INT8)	–	40~60	32MB
ATOM	Rebellions (국내)	FP16, INT8, INT4, INT2	128 (INT8)	32 (FP16)	60~150	64MB

국내외 인공지능 반도체에 대한 향후 연구 방향 제시

- **인공지능 반도체 아키텍처**
  - 특정 계산에 대한 병렬 처리에 최적화 된 아키텍처를 갖춤으로써 많은 연산을 빠르고 효율적으로 수행할 수 있어야 함
  - **특정 데이터 타입과 연산 대상이 되는 텐서의 특징 (Density, Sparse 행렬 등) 에 따라 행렬 곱셈 연산을 가속화 하도록 설계**
    → 딥러닝 연산의 대부분을 차지하는 연산이 가속화 되므로 모델의 성능이 크게 개선
  - 거의 모든 칩에서 통신 과정에서의 오버헤드가 발생하는 경향 존재
  - 향후에는 컴퓨팅 성능 향상, 소비 전력 최소화, 메모리 대역폭 및 통신 오버헤드 관리 간의 균형을 유지하는 데에 초점 두어야 할 것

- **소프트웨어 및 알고리즘 최적화**
  - 소프트웨어 및 알고리즘을 최적화하여 인공지능 반도체의 효율성을 향상시킬 수 있음
  - **소프트웨어 최적화는 하드웨어의 성능을 최대화하기 위한 소프트웨어 및 알고리즘의 최적화를 의미**
  - 특히 앞서 언급한 pruning, quantization, zero value skipping 등과 같은 기술을 통해 **계산 부하를 줄이고 메모리 사용량과 에너지 소비를 최소화**
  - 인공지능 작업을 위한 **명령어 세트 아키텍처 (Instruction Set Architecture, ISA)에 대한 연구도 필요**할 듯 함
    - 딥러닝 연산을 제공하는데 **과도한 하드웨어 리소스를 사용**하도록 하는 경우가 있으므로 명령어 수준에서의 유연성을 확보할 필요가 있음
    - 각 모델에 대한 명령어가 아니라 **여러 모델에 대해 범용적**으로 사용될 수 있다면 더 좋을 것
  - 이처럼 인공지능 반도체의 효율성 및 고속화를 위해서는 하드웨어 뿐만 아니라 효율적인 소프트웨어 요소들이 뒷받침되어야 함!

- **에너지 효율성 및 고속화를 위한 향후 연구 방향 제시** (위의 두 카테고리를 고려)
  - 딥러닝 알고리즘과 인공지능 반도체 설계의 상호 작용의 극대화함으로써 성능 향상 (현재 진행 중이며 지속되어야 함)
  - 성능 향상과 더불어 이제는 **에너지를 효율적으로 소모하기 위한 여러 연구들이 필요**
    → 현재 대부분의 인공지능 반도체들이 높은 성능을 목표로 함 → **활용 목적에 맞는 에너지 효율성**을 갖춰나가야 할 것
  - **하드웨어 유연성 확보**
    - 다양한 딥러닝 작업에 사용할 수 있는 유연성을 갖춘 반도체의 개발과 소프트웨어와의 호환성이 중요
    - 레이어 별 연산 제공 및 사용 시간에 따른 최적 연산 재구성 등이 가능한 아키텍처에 대한 연구가 필요할 것
  - 인공지능 작업의 특성상 개인 정보와 같은 민감 데이터를 처리하는 데에 널리 활용되고 있으므로, **보안 및 프라이버시 보호 기능을 내장한 설계 또한 중요**

결론

- 국내외 인공지능 반도체 연구 동향에 대해 분석한 결과:

  단순히 높은 연산 속도만이 중요한 것이 아님

  제공하는 데이터 타입과 전력 소모량 그리고 온 칩 메모리의 크기 등이 전체 연산 효율에 영향을 미침을 확인

  현재로서는 NVDIA의 **H100이 여러 요소에서 가장 선두**

  D1이나 AB9와 같은 칩은 특정 목적에 맞게 설계됨

- 향후에는 **연산 성능 향상과 더불어 목적에 맞는 칩 설계, 에너지 효율성 향상, 하드웨어 및 소프트웨어 간의 최적화 등에 초점을 맞춘 연구가 진행되어야 할 것**으로 생각된다.

암호연구회 사전 작업에 관한 질문

- 메일로 사전 작업한다고 하신 거 지금부터 하면 되나요?
  - 개인에게는 판매하지 않는다.
  - 가격이 아주 아주 비싸다.
  - 위의 답변 듣기 위한 작업

세미나 늦지 않겠습니다.