

100만라인 프로그램 의 전체 분석

오학주
(허기홍, 이원찬, 이우석)

서울대학교 프로그래밍 연구실

ROSAEC Workshop @HKUST
2012.01.18

연구 동기

너무 무거운 스파로우



프로그램	크기(라인수)	시간	메모리
gzip-1.2.4a	7,327	1시간	0.9G
bc-1.06	13,093	7시간	0.8G
make-3.76.1	27,304	24시간	2.7G
bash-2.05a	105,174	n/a	>4G

전체 분석, 3.2GHz with 4GB of memory

정적 분석의 난제

정확하게 (precise),
모든 실행상황을 포섭하며 (sound),
큰 프로그램을 (scalable),
전체 분석 (global analysis) 하기

현실

안전성 또는 큰 프로그램을 포기

“bug-finders”

scalable
unsound

“verifiers”

sound
unscalable

스패로우의 경우

- 전체 분석은 포기
 - 대신 파일별로 분석
 - 정확도가 너무 떨어짐
- 안전성을 포기
 - 파일간 교류는 무시
 - 정확도 회복

목표

의 비용 절감

안전성과 정확성은 유지하면서
전체 분석의 비용을 절감하는 기술 개발

꼭 필요한 일에 집중하기 (Localization)

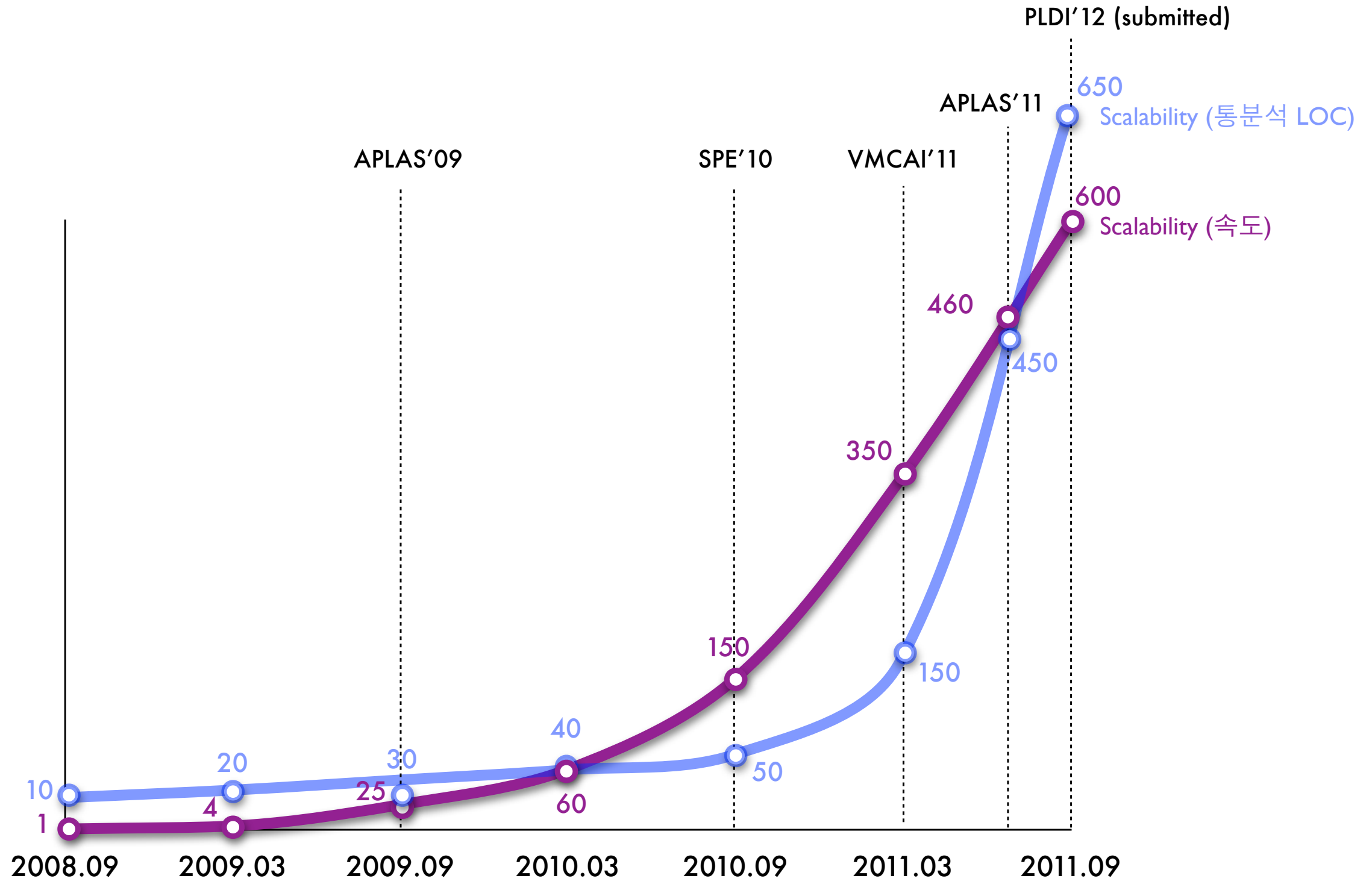
“local reasoning”
“frame rule”

$$\frac{\{P\}C\{Q\}}{\{P*R\}C\{Q*R\}}$$

- 필요한 메모리 영역만 (spatial localization)
- 필요한 시점에만 (temporal localization)
- 필요한 상황에만 (contextual localization)

성능향상 Sparrow

The Early Bird



성능향상



Program	LOC	Baseline		Localize		Spd↑	Mem↓
		Time	Mem	Time	Mem		
gzip-1.2.4a	7 K	772	240	3	63	257 x	74 %
bc-1.06	13 K	1,270	276	7	75	181 x	73 %
less-382	23 K	9,561	1,113	33	127	289 x	86 %
make-3.76.1	27 K	24,240	1,391	21	114	1,154 x	92 %
wget-1.9	35 K	44,092	2,546	11	85	4,008 x	97 %
a2ps-4.14	64 K	∞	N/A	40	353	N/A	N/A
sendmail-8.13.6	130 K	∞	N/A	744	678	N/A	N/A
nethack-3.3.0	211 K	∞	N/A	16,373	5,298	N/A	N/A
emacs-22.1	399 K	∞	N/A	37,830	7,795	N/A	N/A
python-2.5.1	435 K	∞	N/A	11,039	5,535	N/A	N/A
linux-3.0	710 K	∞	N/A	33,618	20,529	N/A	N/A
gimp-2.6	959 K	∞	N/A	3,874	3,602	N/A	N/A
ghostscript-9.00	1,363 K	∞	N/A	14,814	6,384	N/A	N/A

성능향상



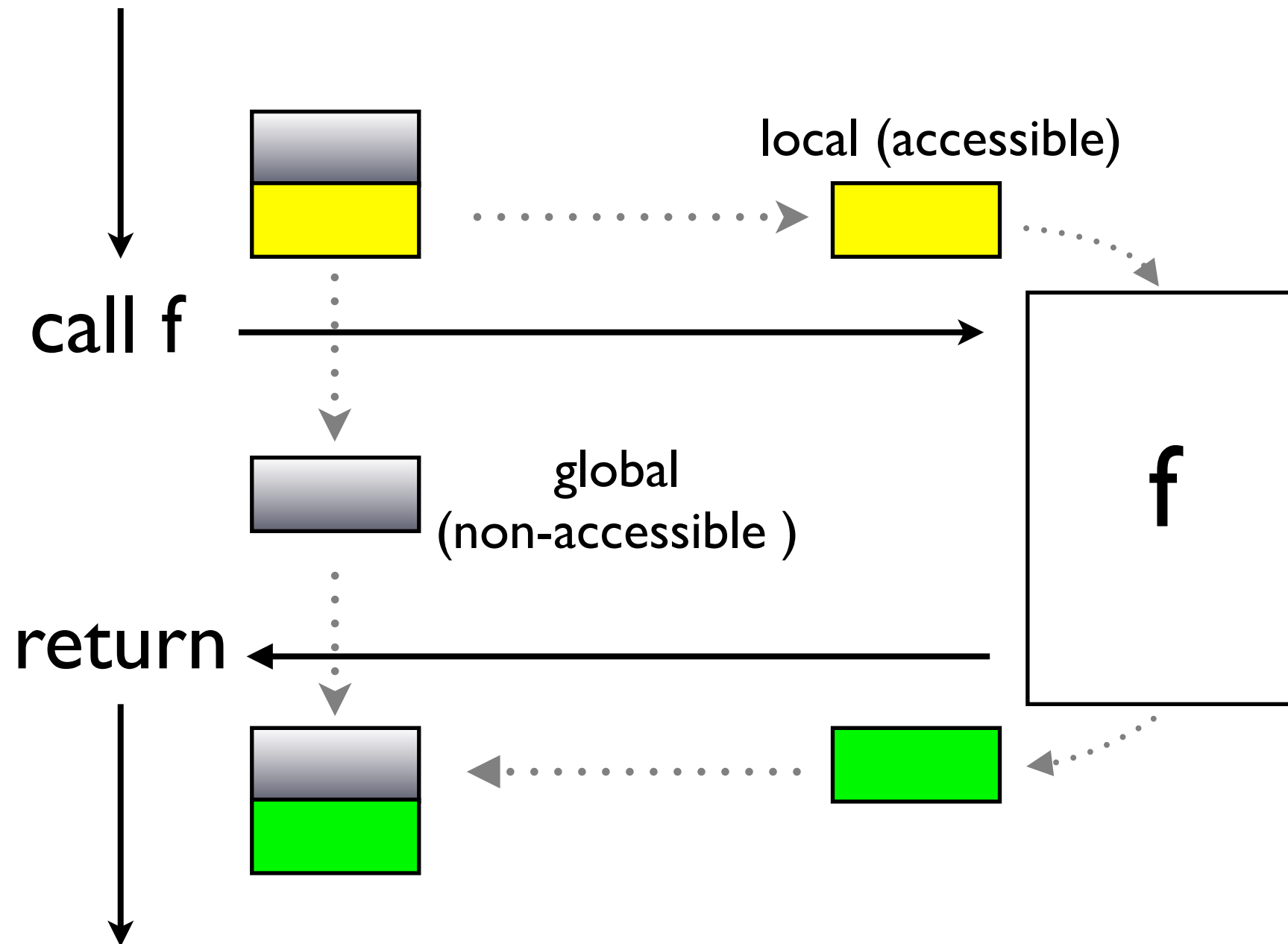
Program	LOC	Baseline		Localize		Spd↑	Mem↓
		Time	Mem	Time	Mem		
gzip-1.2.4a	7 K	772	240	3	63	257 x	74 %
bc-1.06	13 K	1,270	276	7	75	181 x	73 %
less-382	23 K	9,561	1,113	33	127	289 x	86 %
make-3.76.1	27 K	24,240	1,391	21	114	1,154 x	92 %
wget-1.9	35 K	44,092	2,546	11	85	4,008 x	97 %
a2ps-4.14	64 K	∞	N/A	40	353	N/A	N/A
sendmail-8.13.6	130 K	∞	N/A	744	678	N/A	N/A
nethack-3.3.0	211 K	∞	N/A	16,373	5,298	N/A	N/A
emacs-22.1	399 K	∞	N/A	37,830	7,795	N/A	N/A
python-2.5.1	435 K	∞	N/A	11,039	5,535	N/A	N/A
linux-3.0	710 K	∞	N/A	33,618	20,529	N/A	N/A
gimp-2.6	959 K	∞	N/A	3,874	3,602	N/A	N/A
ghostscript-9.00	1,363 K	∞	N/A	14,814	6,384	N/A	N/A

성능향상



Program	LOC	Baseline		Localize		Spd↑	Mem↓
		Time	Mem	Time	Mem		
gzip-1.2.4a	7 K	772	240	3	63	257 x	74 %
bc-1.06	13 K	1,270	276	7	75	181 x	73 %
less-382	23 K	9,561	1,113	33	127	289 x	86 %
make-3.76.1	27 K	24,240	1,391	21	114	1,154 x	92 %
wget-1.9	35 K	44,092	2,546	11	85	4,008 x	97 %
a2ps-4.14	64 K	∞	N/A	40	353	N/A	N/A
sendmail-8.13.6	130 K	∞	N/A	744	678	N/A	N/A
nethack-3.3.0	211 K	∞	N/A	16,373	5,298	N/A	N/A
emacs-22.1	399 K	∞	N/A	37,830	7,795	N/A	N/A
python-2.5.1	435 K	∞	N/A	11,039	5,535	N/A	N/A
linux-3.0	710 K	∞	N/A	33,618	20,529	N/A	N/A
gimp-2.6	959 K	∞	N/A	3,874	3,602	N/A	N/A
ghostscript-9.00	1,363 K	∞	N/A	14,814	6,384	N/A	N/A

필요한 메모리 영역 잘라내기 (spatial localization)



효과

```
int g;
```

```
int f() {...}
```

f does not access g

```
int main() {
```

```
    g = 0;
```

```
    f();
```

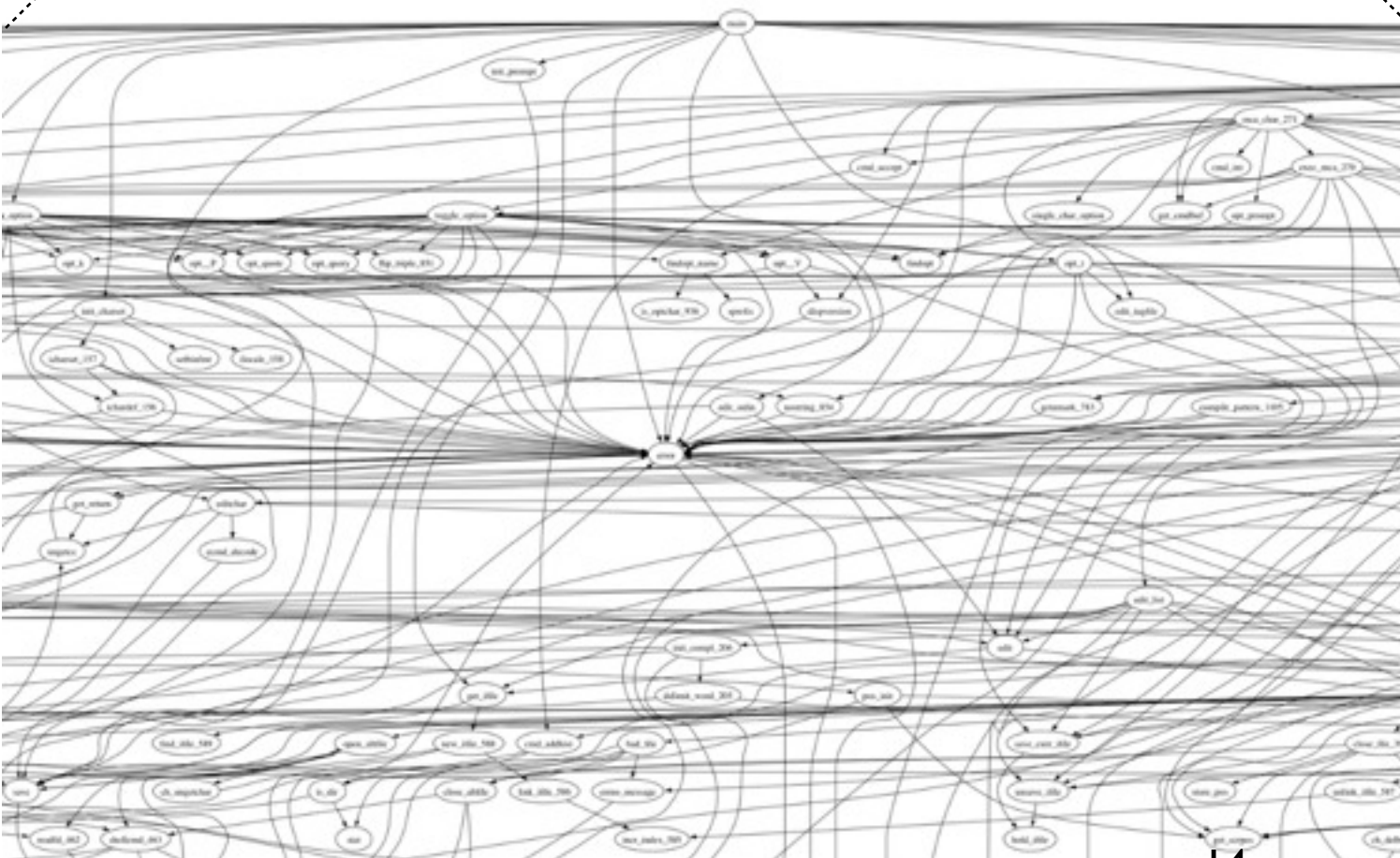
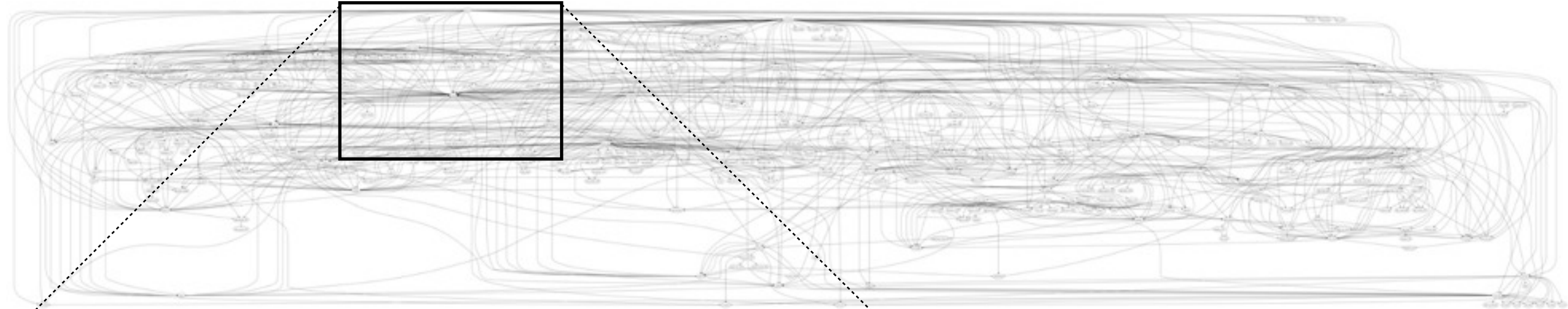
```
    g = 1;
```

```
    f();
```

```
}
```

실제 프로그램 분석에 꼭 필요

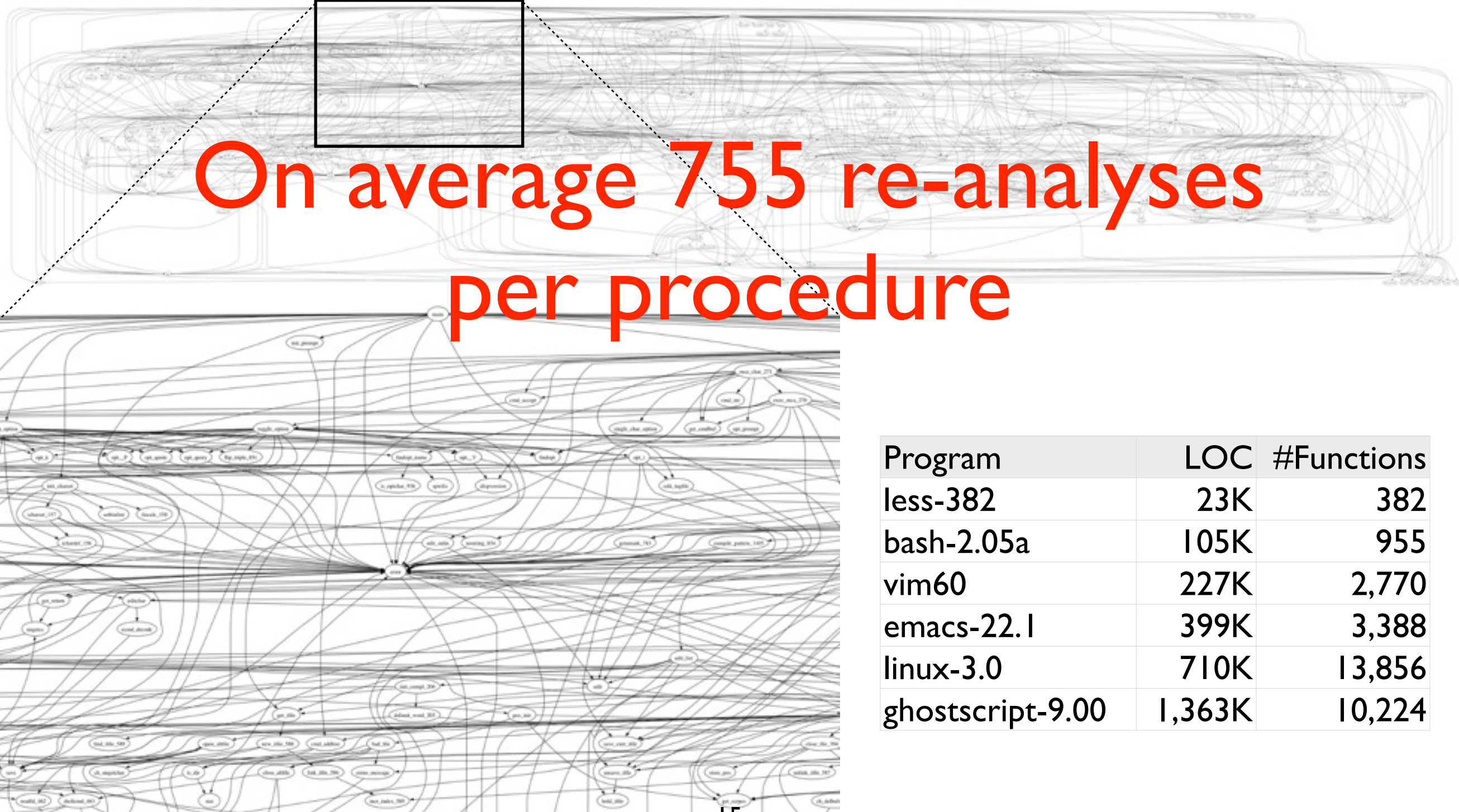
less-382 (23,822 LOC)



Program	LOC	#Functions
less-382	23K	382
bash-2.05a	105K	955
vim60	227K	2,770
emacs-22.1	399K	3,388
linux-3.0	710K	13,856
ghostscript-9.00	1,363K	10,224

실제 프로그램 분석에 꼭 필요

less-382 (23,822 LOC)

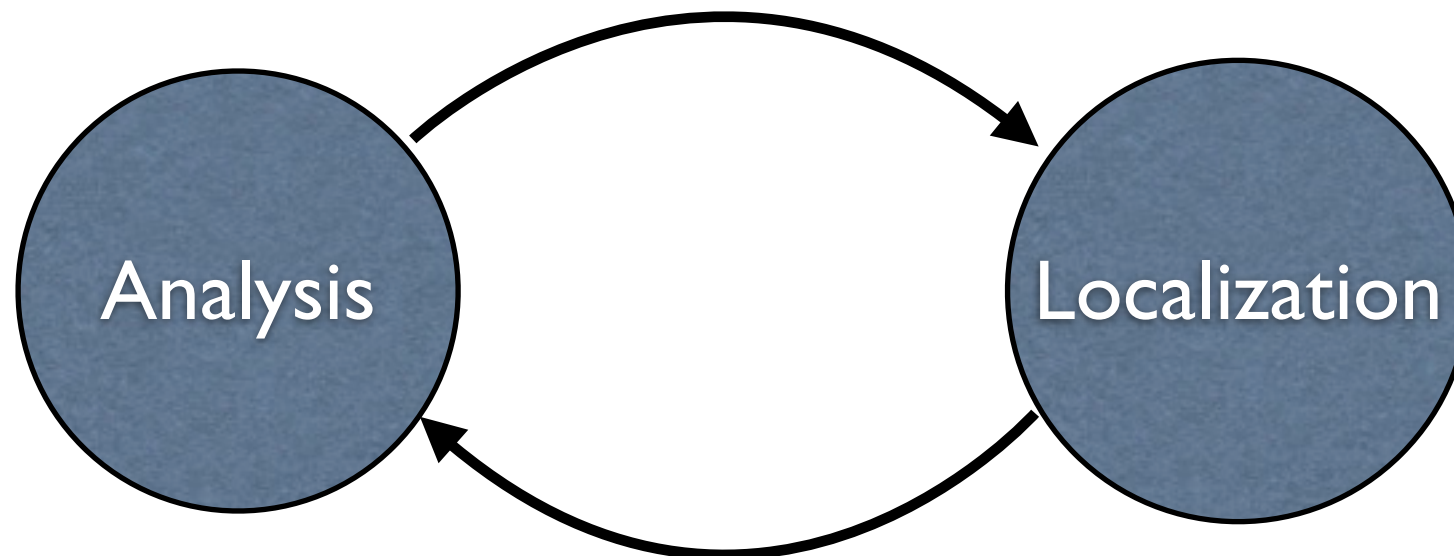


On average 755 re-analyses
per procedure

Program	LOC	#Functions
less-382	23K	382
bash-2.05a	105K	955
vim60	227K	2,770
emacs-22.1	399K	3,388
linux-3.0	710K	13,856
ghostscript-9.00	1,363K	10,224

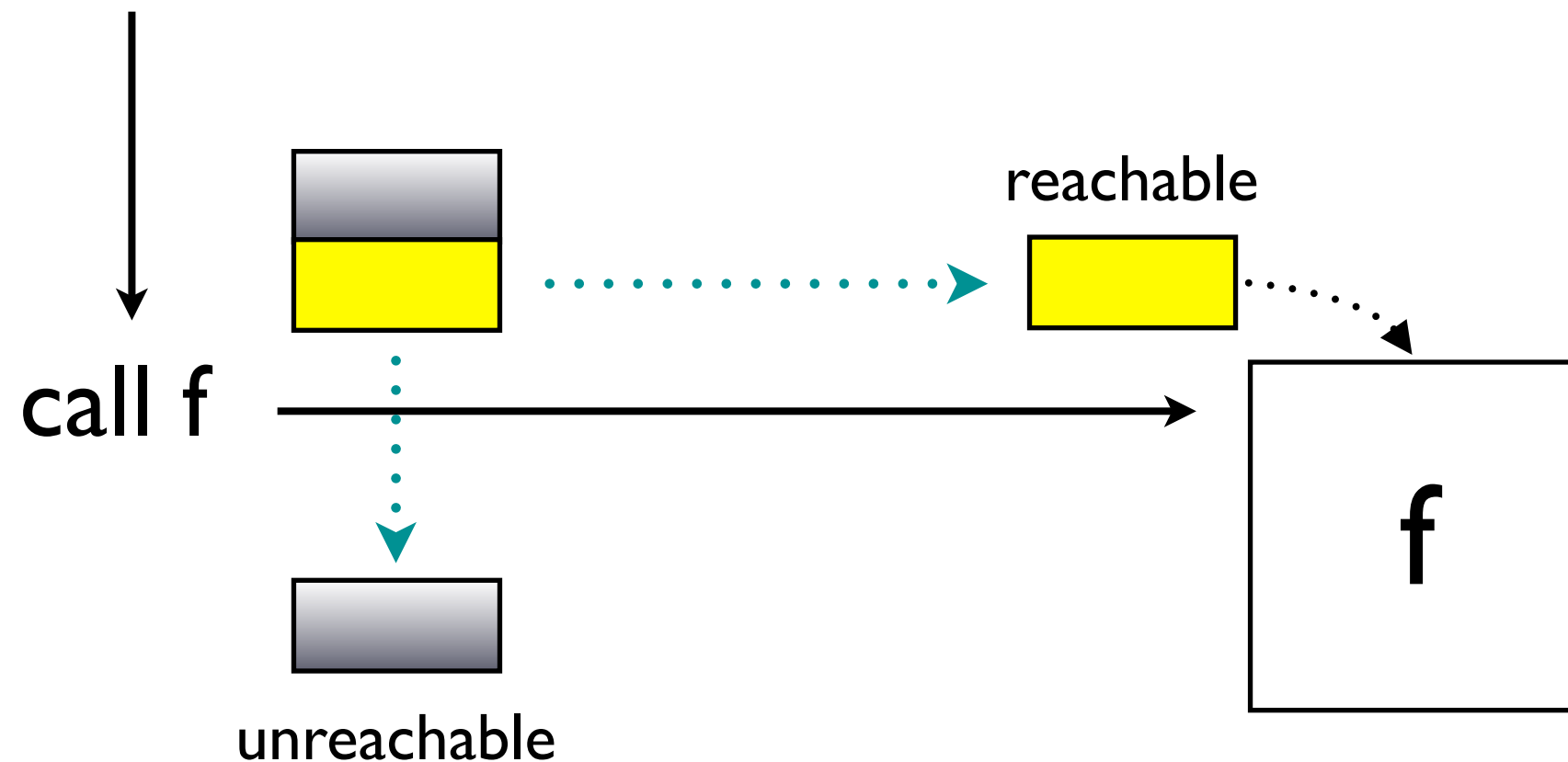
어려움

필요한 메모리 영역은 분석해봐야 알 수 있음



접근가능한 메모리 영역만 넘기기 (Reachability-based Localization / abstract garbage collection)

- 전역 주소, 함수 인자로부터 접근 가능한 영역

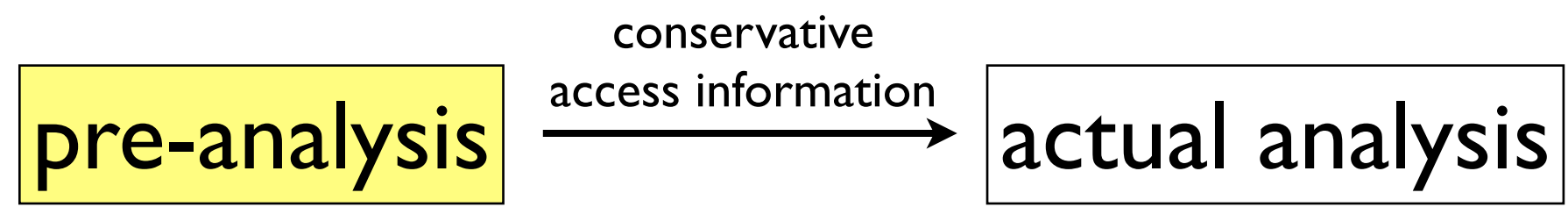


너무 조심스런 방식 (too conservative)

Program	LOC	accessed memory / reachable memory
spell-1.0	2,213	5 / 453 (1.1%)
barcode-0.96	4,460	19 / 1175 (1.6%)
httptunnel-3.3	6,174	10 / 673 (1.5%)
gzip-1.2.4a	7,327	22 / 1002 (2.2%)
jwhois-3.0.1	9,344	28 / 830 (3.4%)
parser	10,900	75 / 1787 (4.2%)
bc-1.06	13,093	24 / 824 (2.9%)
less-290	18,449	86 / 1546 (5.6%)

average : 4%

실제로 쓰는 영역만 자르기 (access-based localization)

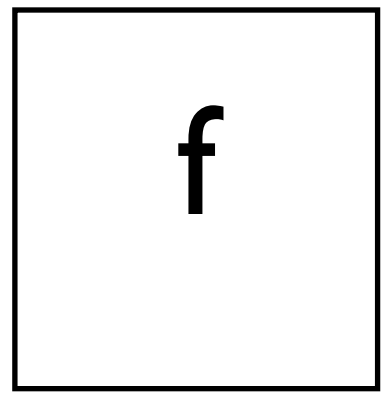


Over-approximation of actual access info.

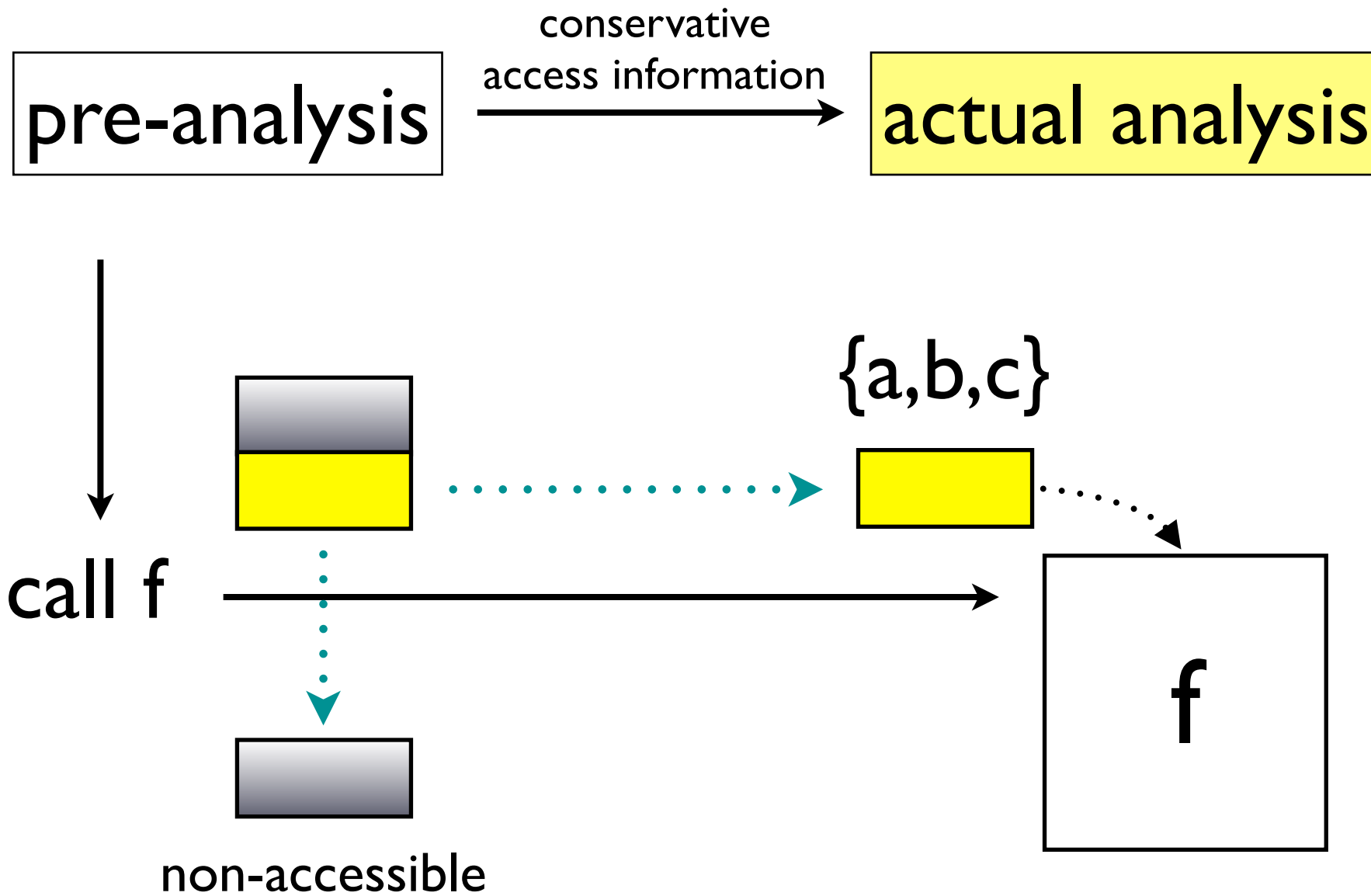
{a,b,c}
U

actual access info.

{a,b}

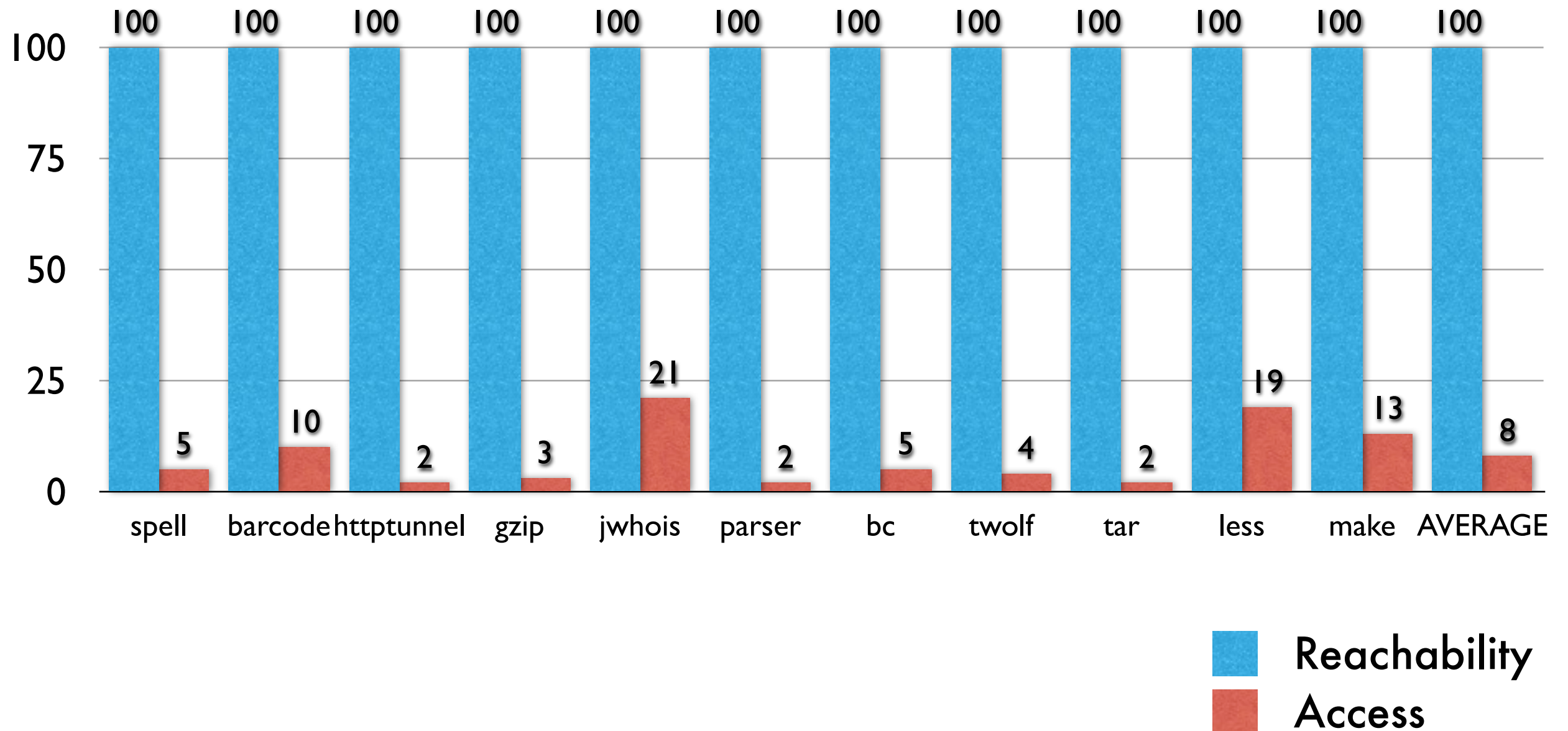


실제로 쓰는 영역만 자르기 (access-based localization)

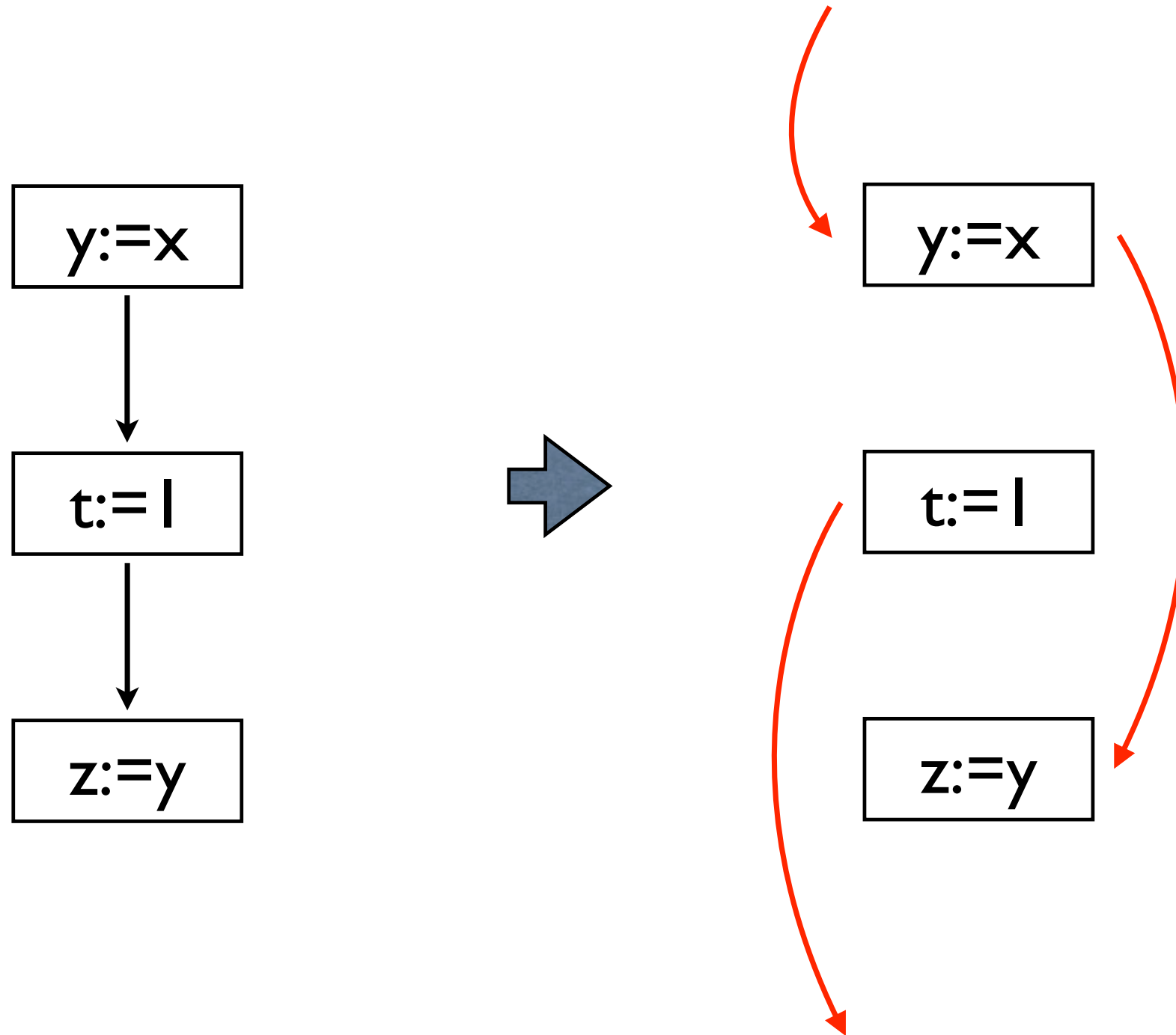


성능 향상

5x~50x speed-up over reachability

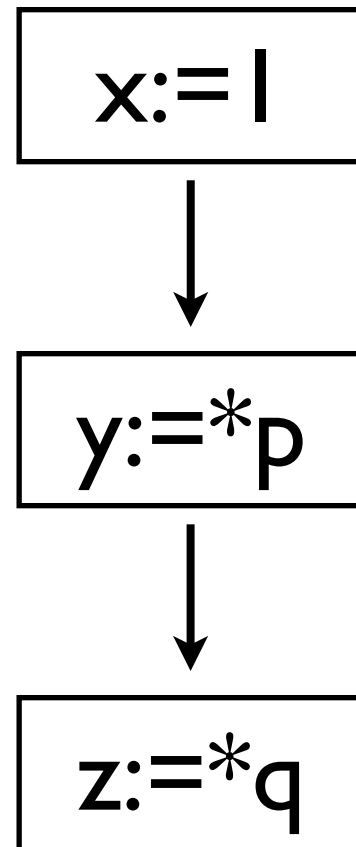


필요한 시점에만 분석하기 (temporal localization / sparse analysis)

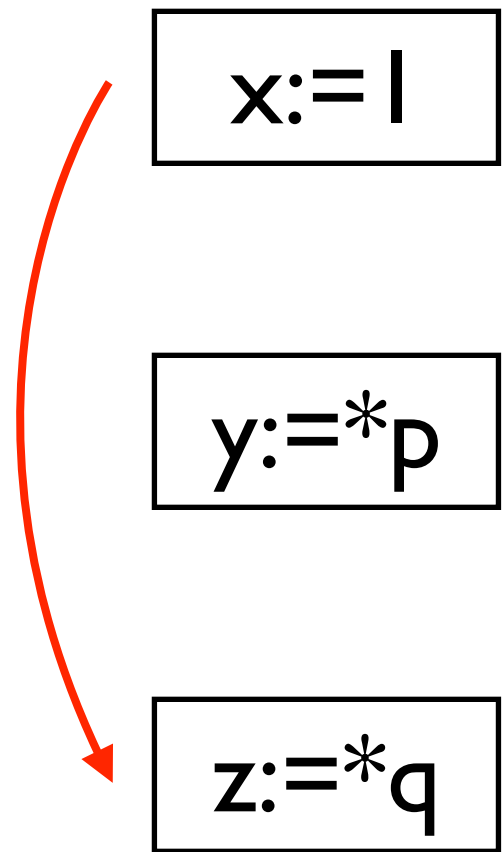


어려움 (1)

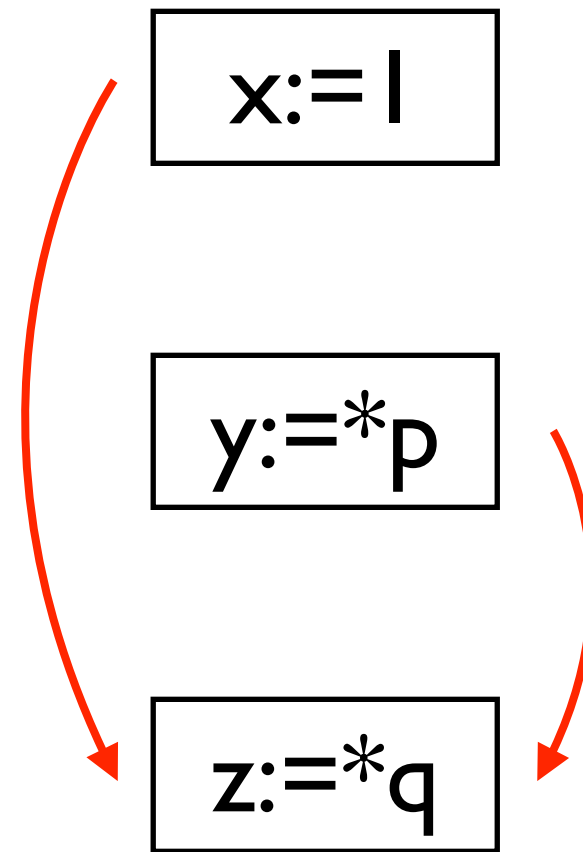
어떤값이 어디서 필요한지 미리 알 수 없음



안전하게만 가능

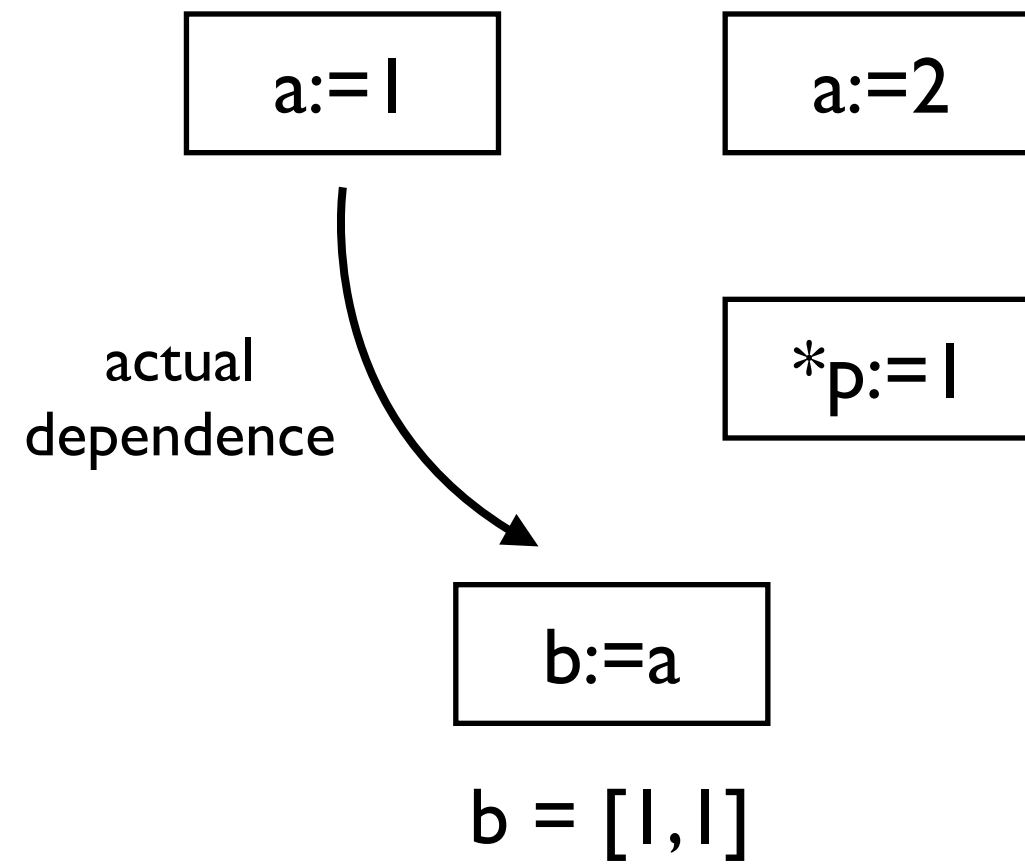


actual

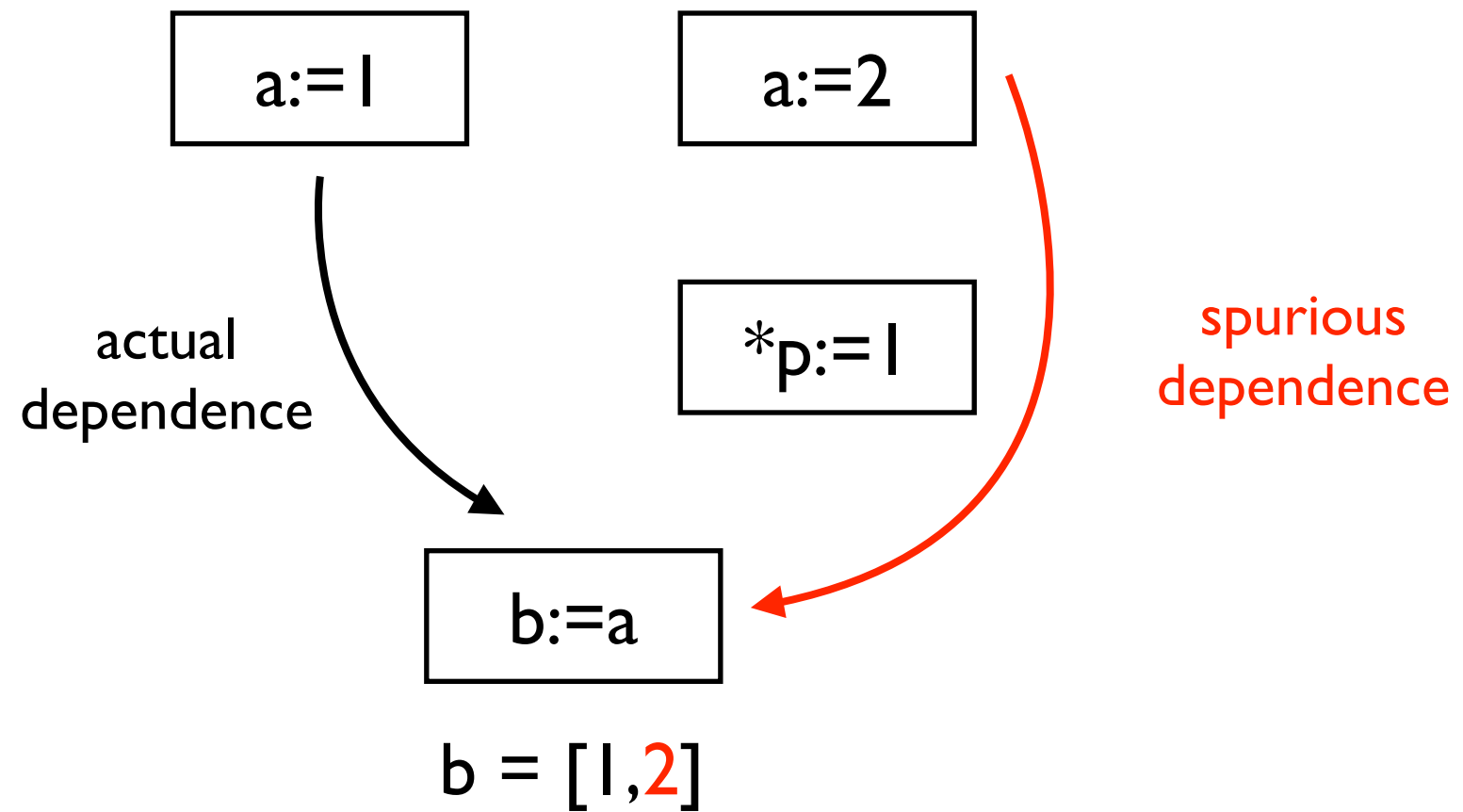


conservative

어려움 (2)



어려움 (2)

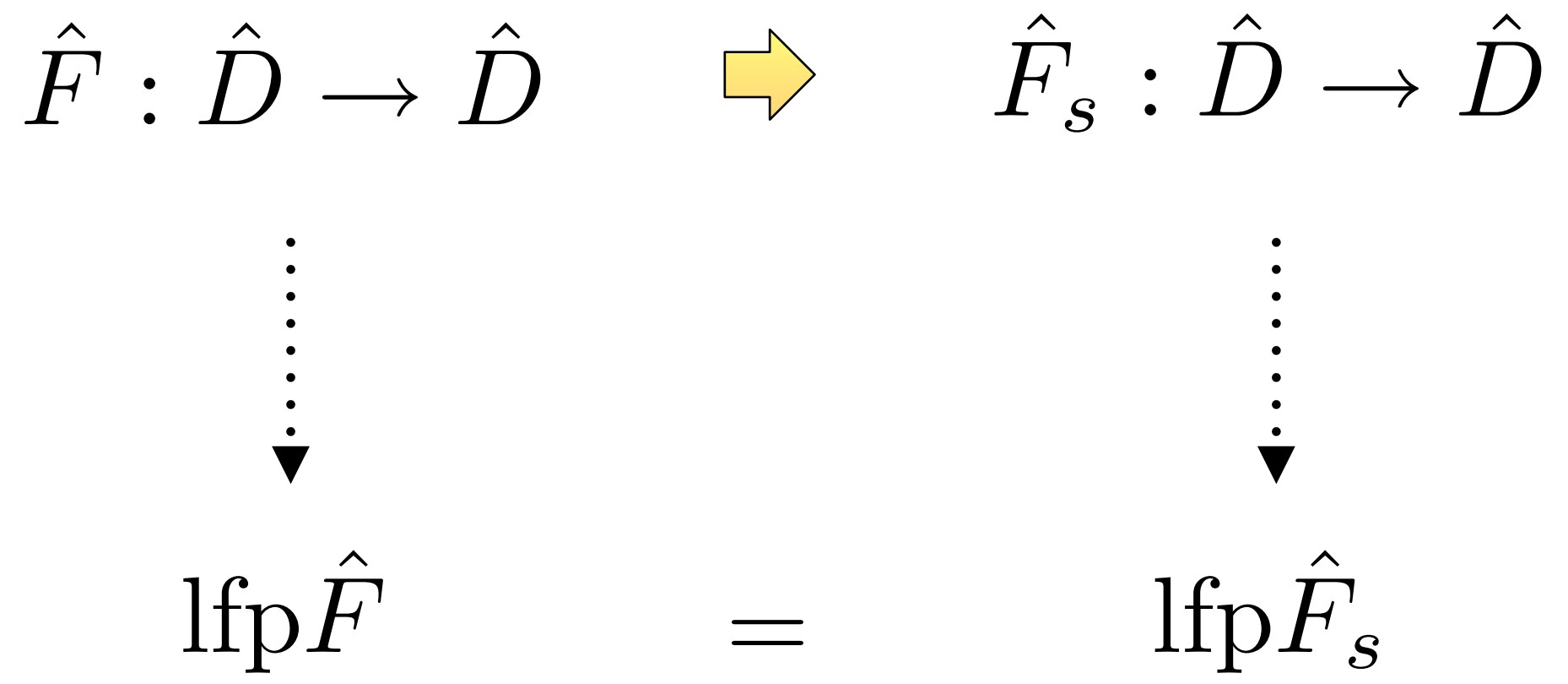


관련 연구

- 특정 분석에 제한된 “알고리즘”들만 존재
- 주로 포인터 분석 성능향상을 위한

틀 (framework)

일반적인 요약해석에 대해,



성능 향상: 값 분석 (non-relational analysis)

Programs	LOC	Interval _{vanilla}		Interval _{base}		Spd _{↑1}	Mem _{↓1}	Interval _{sparse}		Spd _{↑2}	Mem _{↓2}
		Time	Mem	Time	Mem			Time	Mem		
gzip-1.2.4a	7K	772	240	14	65	55 x	73 %	3	63	5 x	3 %
bc-1.06	13K	1,270	276	96	126	13 x	54 %	7	75	14 x	40 %
tar-1.13	20K	12,947	881	338	177	38 x	80 %	8	93	42 x	47 %
less-382	23K	9,561	1,113	1,211	378	8 x	66 %	33	127	37 x	66 %
make-3.76.1	27K	24,240	1,391	1,893	443	13 x	68 %	21	114	90 x	74 %
wget-1.9	35K	44,092	2,546	1,214	378	36 x	85 %	11	85	110 x	78 %
screen-4.0.2	45K	∞	N/A	31,324	3,996	N/A	N/A	767	303	41 x	92 %
a2ps-4.14	64K	∞	N/A	3,200	1,392	N/A	N/A	40	353	80 x	75 %
bash-2.05a	105K	∞	N/A	1,683	1,386	N/A	N/A	67	220	25 x	84 %
lsh-2.0.4	111K	∞	N/A	45,522	5,266	N/A	N/A	471	577	97 x	89 %
sendmail-8.13.6	130K	∞	N/A	∞	N/A	N/A	N/A	744	678	N/A	N/A
nethack-3.3.0	211K	∞	N/A	∞	N/A	N/A	N/A	16,373	5,298	N/A	N/A
vim60	227K	∞	N/A	∞	N/A	N/A	N/A	23,798	5,190	N/A	N/A
emacs-22.1	399K	∞	N/A	∞	N/A	N/A	N/A	37,830	7,795	N/A	N/A
python-2.5.1	435K	∞	N/A	∞	N/A	N/A	N/A	11,039	5,535	N/A	N/A
linux-3.0	710K	∞	N/A	∞	N/A	N/A	N/A	33,618	20,529	N/A	N/A
gimp-2.6	959K	∞	N/A	∞	N/A	N/A	N/A	3,874	3,602	N/A	N/A
ghostscript-9.00	1,363K	∞	N/A	∞	N/A	N/A	N/A	14,814	6,384	N/A	N/A

성능 향상: 값 분석 (non-relational analysis)

Programs	LOC	Interval _{vanilla}		Interval _{base}		Spd _{↑1}	Mem _{↓1}	Interval _{sparse}		Spd _{↑2}	Mem _{↓2}
		Time	Mem	Time	Mem			Time	Mem		
gzip-1.2.4a	7K	772	240	14	65	55 x	73 %	3	63	5 x	3 %
bc-1.06	13K	1,270	276	96	126	13 x	54 %	7	75	14 x	40 %
tar-1.13	20K	12,947	881	338	177	38 x	80 %	8	93	42 x	47 %
less-382	23K	9,561	1,113	1,211	378	8 x	66 %	33	127	37 x	66 %
make-3.76.1	27K	24,240	1,391	1,893	443	13 x	68 %	21	114	90 x	74 %
wget-1.9	35K	44,092	2,546	1,214	378	36 x	85 %	11	85	110 x	78 %
screen-4.0.2	45K	∞	N/A	31,324	3,996	N/A	N/A	767	303	41 x	92 %
a2ps-4.14	64K	∞	N/A	3,200	1,392	N/A	N/A	40	353	80 x	75 %
bash-2.05a	105K	∞	N/A	1,683	1,386	N/A	N/A	67	220	25 x	84 %
lsh-2.0.4	111K	∞	N/A	45,522	5,266	N/A	N/A	471	577	97 x	89 %
sendmail-8.13.6	130K	∞	N/A	∞	N/A	N/A	N/A	744	678	N/A	N/A
nethack-3.3.0	211K	∞	N/A	∞	N/A	N/A	N/A	16,373	5,298	N/A	N/A
vim60	227K	∞	N/A	∞	N/A	N/A	N/A	23,798	5,190	N/A	N/A
emacs-22.1	399K	∞	N/A	∞	N/A	N/A	N/A	37,830	7,795	N/A	N/A
python-2.5.1	435K	∞	N/A	∞	N/A	N/A	N/A	11,039	5,535	N/A	N/A
linux-3.0	710K	∞	N/A	∞	N/A	N/A	N/A	33,618	20,529	N/A	N/A
gimp-2.6	959K	∞	N/A	∞	N/A	N/A	N/A	3,874	3,602	N/A	N/A
ghostscript-9.00	1,363K	∞	N/A	∞	N/A	N/A	N/A	14,814	6,384	N/A	N/A

성능 향상: 값 분석 (non-relational analysis)

Programs	LOC	Interval _{vanilla}		Interval _{base}		Spd _{↑1}	Mem _{↓1}	Interval _{sparse}		Spd _{↑2}	Mem _{↓2}
		Time	Mem	Time	Mem			Time	Mem		
gzip-1.2.4a	7K	772	240	14	65	55 x	73 %	3	63	5 x	3 %
bc-1.06	13K	1,270	276	96	126	13 x	54 %	7	75	14 x	40 %
tar-1.13	20K	12,947	881	338	177	38 x	80 %	8	93	42 x	47 %
less-382	23K	9,561	1,113	1,211	378	8 x	66 %	33	127	37 x	66 %
make-3.76.1	27K	24,240	1,391	1,893	443	13 x	68 %	21	114	90 x	74 %
wget-1.9	35K	44,092	2,546	1,214	378	36 x	85 %	11	85	110 x	78 %
screen-4.0.2	45K	∞	N/A	31,324	3,996	N/A	N/A	767	303	41 x	92 %
a2ps-4.14	64K	∞	N/A	3,200	1,392	N/A	N/A	40	353	80 x	75 %
bash-2.05a	105K	∞	N/A	1,683	1,386	N/A	N/A	67	220	25 x	84 %
lsh-2.0.4	111K	∞	N/A	45,522	5,266	N/A	N/A	471	577	97 x	89 %
sendmail-8.13.6	130K	∞	N/A	∞	N/A	N/A	N/A	744	678	N/A	N/A
nethack-3.3.0	211K	∞	N/A	∞	N/A	N/A	N/A	16,373	5,298	N/A	N/A
vim60	227K	∞	N/A	∞	N/A	N/A	N/A	23,798	5,190	N/A	N/A
emacs-22.1	399K	∞	N/A	∞	N/A	N/A	N/A	37,830	7,795	N/A	N/A
python-2.5.1	435K	∞	N/A	∞	N/A	N/A	N/A	11,039	5,535	N/A	N/A
linux-3.0	710K	∞	N/A	∞	N/A	N/A	N/A	33,618	20,529	N/A	N/A
gimp-2.6	959K	∞	N/A	∞	N/A	N/A	N/A	3,874	3,602	N/A	N/A
ghostscript-9.00	1,363K	∞	N/A	∞	N/A	N/A	N/A	14,814	6,384	N/A	N/A

성능 향상: 관계 분석 (relational analysis)

Programs	Octagon _{vanilla}		Octagon _{base}		Spd _{↑1}	Mem _{↓1}	Octagon _{sparse}		Spd _{↑2}	Mem _{↓2}
	Time	Mem	Time	Mem			Total	Mem		
gzip-1.2.4a	9,649	5,744	483	1,355	20 x	76 %	15	211	30 x	84 %
bc-1.06	15,027	10,090	1,454	5,065	10 x	50 %	21	381	55 x	92 %
tar-1.13	∞	N/A	21,125	13,810	N/A	N/A	52	588	377 x	95 %
less-382	∞	N/A	∞	N/A	N/A	N/A	131	405	N/A	N/A
make-3.76.1	∞	N/A	∞	N/A	N/A	N/A	39	554	N/A	N/A
wget-1.9	∞	N/A	∞	N/A	N/A	N/A	189	1,098	N/A	N/A
screen-4.0.2	∞	N/A	∞	N/A	N/A	N/A	7,169	19,143	N/A	N/A
a2ps-4.14	∞	N/A	∞	N/A	N/A	N/A	794	1,229	N/A	N/A
bash-2.05a	∞	N/A	∞	N/A	N/A	N/A	407	1,875	N/A	N/A
lsh-2.0.4	∞	N/A	∞	N/A	N/A	N/A	1,553	4,449	N/A	N/A
sendmail-8.13.6	∞	N/A	∞	N/A	N/A	N/A	1,463	9,881	N/A	N/A

결론

큰 프로그램의 의미기반 정적분석이 가능