

Conociendo a Eulogia

Carlos Bederián, CCAD-UNC

carlos.bederian@unc.edu.ar



CCAD

Centro de
Computación
de Alto
Desempeño



UNC

Universidad
Nacional
de Córdoba

Cómo llegamos hasta acá

Cristina

- Propósito general
- Retirada por eficiencia energética



Cómo llegamos hasta acá

Mendieta

- Cluster con GPUs
 - Alta performance/\$
 - Alta performance/W
 - GPGPU no siempre es aplicable



Funding

1. ~300K ARS de UNC para retirar Cristina
2. ~600K ARS subsidio SNCAD y contraparte UNC
3. ~400K ARS PMT FAMAFA y recursos corrientes de CCAD (en curso)

Qué armamos con eso



CCAD

Centro de
Computación
de Alto
Desempeño

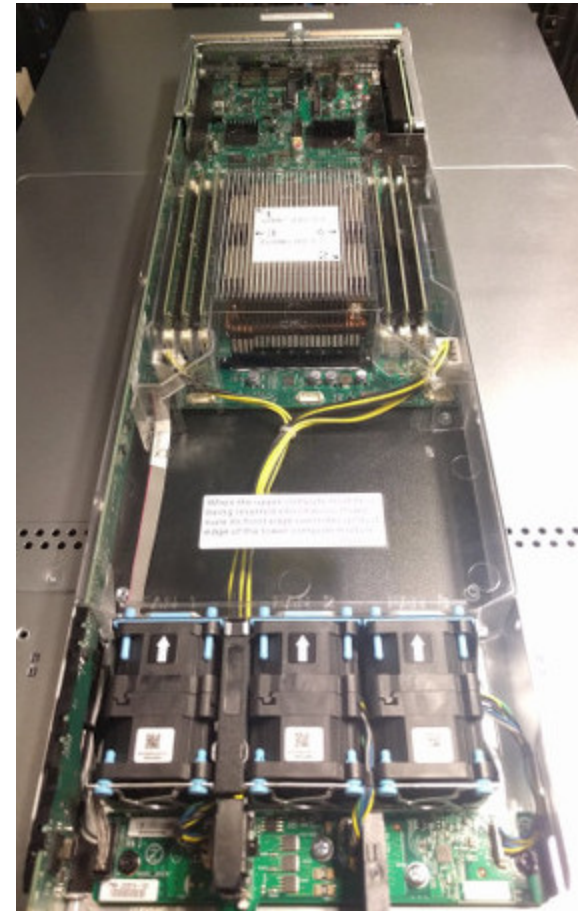


UNC

Universidad
Nacional
de Córdoba

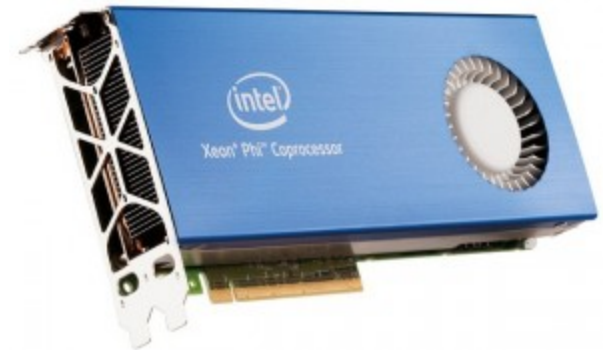
Nodo Eulogia

- Intel Xeon Phi 7210
- 96GB de memoria en 6 canales DDR4-2133
- Disco de estado sólido de 240GB



Larrabee

- Proyecto de Intel para crear una GPU en software
 - Cores x86
 - Memoria coherente
- Nunca vio la luz del día como GPU
- Primer producto: Knights Corner (Xeon Phi x100)
 - Acelerador PCI Express
 - 57-61 cores basados en P54C
 - Unidades de vectores IMCI de 512 bits
 - SMT para ocultar latencia
 - Memoria GDDR5



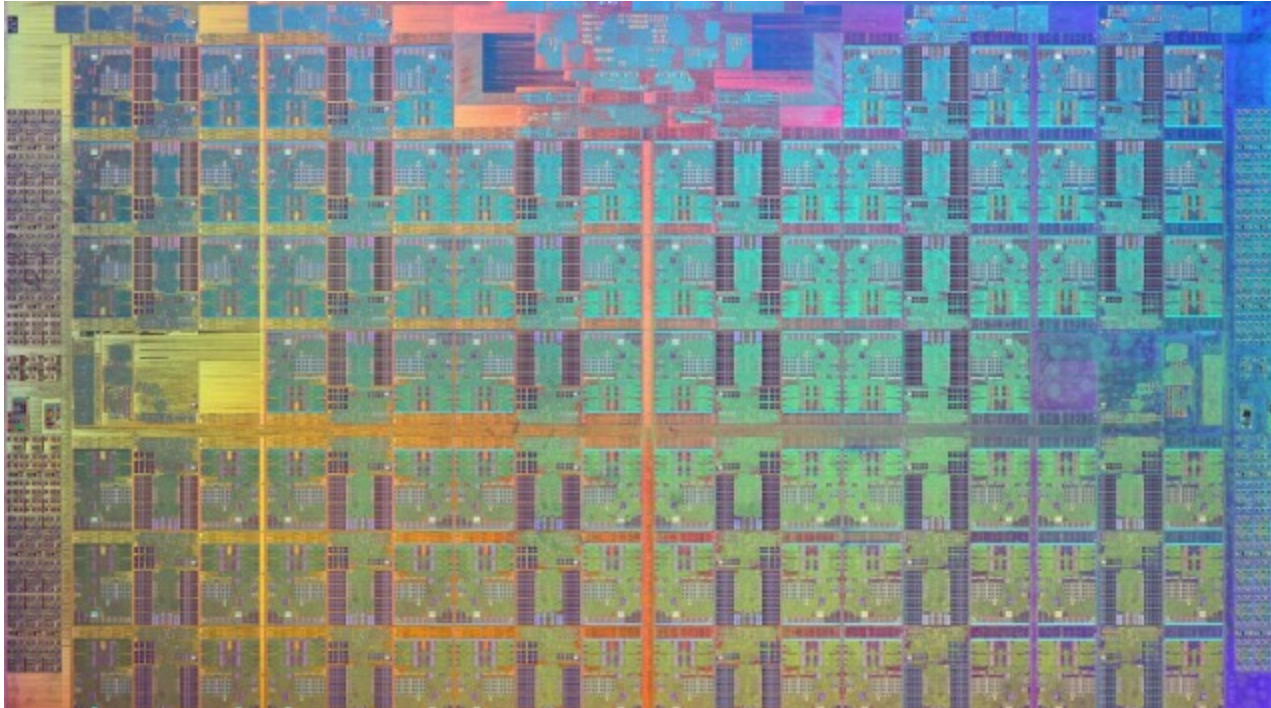
Knights Landing (Xeon Phi x200)

Solución a la mayoría de los problemas de Knights Corner

- Procesador principal, no acelerador
- Soporte X86-64 completo
 - AVX-512, evolución de IMCI que es estándar en adelante
- Memoria expandible

Knights Landing es...lento

- 64 a 72 cores Silvermont (Atom)
 - 1.3 a 1.5 GHz



CCAD

Centro de
Computación
de Alto
Desempeño

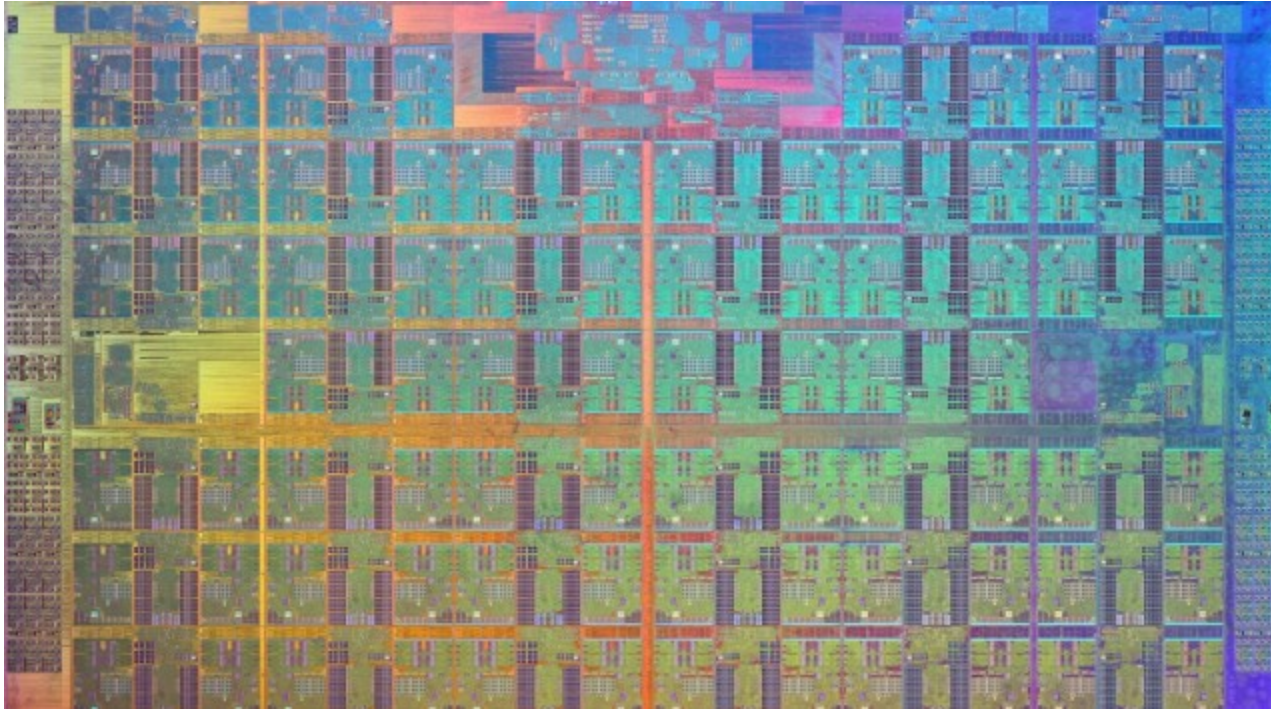


UNC

Universidad
Nacional
de Córdoba

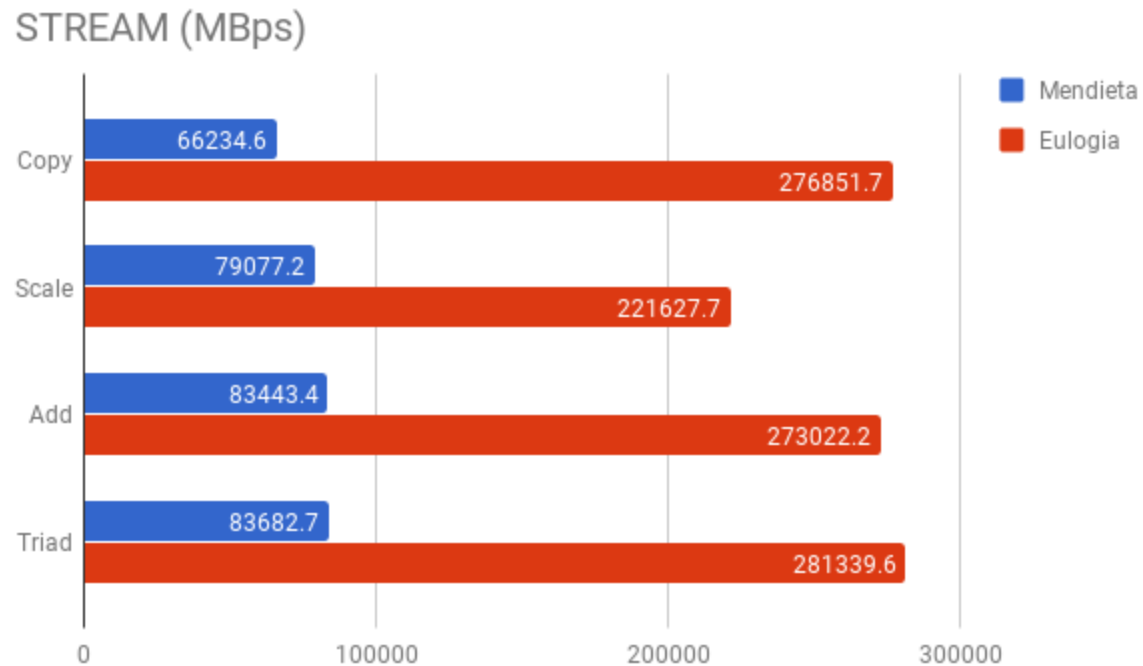
Knights Landing es trabajador

- **Dos** unidades de vectores AVX-512 de 512 bits
- 2.6 a 3.5 TFLOPS FP64



Memoria

- 6 canales de memoria DDR4: ~90GBps
- **16 GB** de memoria HMC integrada: 400-480GBps (pero mayor latencia)
 - Uso como cache o memoria adicional



Comparativa: HPL

- Nodo Mendieta con NVIDIA K20X, ~600W

```
=====
T/V                N    NB    P    Q                Time                Gflops
-----
WR10L2L2          85000  1536    1    2                292.08                1.402e+03
-----
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=          0.0035472 ..... PASSED
=====
```

- Nodo Eulogia, ~350W

```
=====
T/V                N    NB    P    Q                Time                Gflops
-----
WR00C2R2          102000  336    8    8                524.66                1.348e+03
-----
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=          0.0012399 ..... PASSED
=====
```

Comparativa: QE

Benchmark AUSURF112

- Nodo Mendieta, QE 5.4.0, 20 procesos, sin GPUs, ~300W

PWSCF : 56m15.08s CPU 57m11.79s WALL

- Nodo Eulogia, QE 6.2.1, 64 procesos, ~350W

PWSCF : 20m 2.92s CPU 20m46.55s WALL

Caveat emptor

- Resultados preliminares
- Hyperthreading empeora los resultados en ambos casos
 - Quantum Espresso y HPL hacen uso intensivo de MKL
 - No es el caso para todos los programas



CCAD

Centro de
Computación
de Alto
Desempeño



UNC

Universidad
Nacional
de Córdoba

Estado actual del cluster

- 11 nodos de cómputo, 1 actuando como cabecera
 - Adquisición de nodo cabecera en curso
- Storage compartido con el resto de los equipos del CCAD
 - Adquisición de dos servidores de storage en curso
- Instalación Infiniband en curso
 - Trabajos multi nodo
 - Mejoras en acceso a almacenamiento
 - Gracias a Intel y McAfee por sus donaciones

Aplicaciones

Software soportado actualmente

- Intel Parallel Studio XE 2018 (`icc` e `ifort`)
- Intel MKL
- OpenMPI 3.x
- Quantum Espresso

Toolchains en consideración

- GCC 7.x (8.x?)
- LLVM: clang y flang

Más aplicaciones

- Solicitarlas a **soporte**
- Mejor tiempo de respuesta con la llegada de nueva cabecera



CCAD

Centro de
Computación
de Alto
Desempeño



UNC

Universidad
Nacional
de Córdoba

Flags de compiladores

- Es imprescindible activar AVX-512 para obtener buena performance
- La nueva cabecera no soporta el mismo subconjunto de AVX-512 que los nodos
 - Intel: `-xCOMMON-AVX512 -axMIC-AVX512` en vez de `-xHost`
 - GCC: `-march=knl -mno-avx512pf -mno-avx512er` en vez de `--march=native`

Ante la duda revisar los CFLAGS del módulo del compilador

Lanzamiento de trabajos

Hay 64 cores que ocupar, con hasta 4 hilos cada uno

- `--ntasks * cpus-per-task` deberían ser 64, 128, 192 o 256
- La configuración óptima depende del programa y el problema

Recordatorio:

- `#SBATCH --ntasks`: Cantidad de procesos (e.g. MPI) a lanzar
- `#SBATCH --cpus-per-task`: Cantidad de hilos a reservar por proceso
 - Nota: A los hilos los lanza el programa (e.g. `OMP_NUM_THREADS`), SLURM sólo los reserva

Numeracion de cores

- Los procesadores i , $i+64$, $i+128$, $i+192$ de Linux corresponden al mismo core
 - Ojo con la distribución de procesos e hilos!
- Cada fila de htop corresponde a los 4 hilos de un mismo core

1	[100.0%]	65	[0.6%]	129	[0.0%]	193	[0.0%]
2	[100.0%]	66	[0.0%]	130	[0.0%]	194	[0.0%]
3	[100.0%]	67	[0.0%]	131	[0.0%]	195	[0.0%]
4	[100.0%]	68	[0.0%]	132	[0.0%]	196	[0.0%]
5	[100.0%]	69	[0.0%]	133	[0.0%]	197	[0.0%]
6	[100.0%]	70	[0.0%]	134	[0.0%]	198	[0.0%]
7	[100.0%]	71	[1.8%]	135	[0.0%]	199	[0.0%]
8	[100.0%]	72	[0.0%]	136	[0.0%]	200	[0.0%]
9	[100.0%]	73	[2.3%]	137	[0.0%]	201	[0.0%]
10	[100.0%]	74	[0.0%]	138	[0.0%]	202	[0.0%]
11	[100.0%]	75	[0.9%]	139	[0.0%]	203	[0.0%]
12	[100.0%]	76	[0.0%]	140	[0.0%]	204	[0.0%]
13	[100.0%]	77	[1.4%]	141	[0.0%]	205	[0.0%]
14	[100.0%]	78	[0.0%]	142	[0.0%]	206	[0.0%]
15	[100.0%]	79	[0.5%]	143	[0.0%]	207	[0.0%]
16	[100.0%]	80	[0.0%]	144	[0.0%]	208	[0.0%]
17	[100.0%]	81	[0.5%]	145	[0.0%]	209	[0.0%]
18	[100.0%]	82	[0.0%]	146	[0.0%]	210	[0.0%]
19	[100.0%]	83	[0.0%]	147	[0.0%]	211	[0.0%]
20	[100.0%]	84	[0.0%]	148	[0.0%]	212	[0.0%]
21	[100.0%]	85	[3.6%]	149	[0.0%]	213	[0.0%]
22	[100.0%]	86	[0.0%]	150	[0.0%]	214	[0.0%]
23	[100.0%]	87	[0.5%]	151	[0.0%]	215	[0.0%]
24	[100.0%]	88	[0.0%]	152	[0.0%]	216	[0.0%]
25	[100.0%]	89	[0.0%]	153	[0.0%]	217	[0.0%]
26	[100.0%]	90	[0.0%]	154	[0.0%]	218	[0.0%]
27	[100.0%]	91	[0.5%]	155	[0.0%]	219	[0.0%]
28	[100.0%]	92	[0.0%]	156	[0.0%]	220	[0.0%]
29	[100.0%]	93	[0.5%]	157	[0.0%]	221	[0.0%]
30	[100.0%]	94	[0.0%]	158	[0.0%]	222	[0.0%]
31	[100.0%]	95	[1.8%]	159	[0.0%]	223	[0.0%]
32	[100.0%]	96	[0.0%]	160	[0.0%]	224	[0.0%]
33	[100.0%]	97	[0.5%]	161	[0.0%]	225	[0.0%]
34	[100.0%]	98	[0.0%]	162	[0.0%]	226	[0.0%]
35	[100.0%]	99	[1.4%]	163	[0.0%]	227	[0.0%]
36	[100.0%]	100	[0.0%]	164	[0.0%]	228	[0.0%]
37	[100.0%]	101	[2.7%]	165	[0.0%]	229	[0.0%]
38	[100.0%]	102	[0.0%]	166	[0.0%]	230	[0.0%]
39	[100.0%]	103	[0.9%]	167	[0.0%]	231	[0.0%]
40	[100.0%]	104	[0.0%]	168	[0.0%]	232	[0.0%]
41	[100.0%]	105	[0.5%]	169	[0.0%]	233	[0.0%]
42	[100.0%]	106	[0.0%]	170	[0.0%]	234	[0.0%]
43	[100.0%]	107	[1.8%]	171	[0.0%]	235	[0.0%]
44	[100.0%]	108	[0.0%]	172	[0.0%]	236	[0.0%]
45	[100.0%]	109	[0.5%]	173	[0.0%]	237	[0.0%]
46	[100.0%]	110	[0.0%]	174	[0.0%]	238	[0.0%]
47	[100.0%]	111	[0.5%]	175	[0.0%]	239	[0.0%]
48	[100.0%]	112	[0.0%]	176	[0.0%]	240	[0.0%]
49	[100.0%]	113	[0.5%]	177	[0.0%]	241	[0.0%]
50	[100.0%]	114	[0.0%]	178	[0.0%]	242	[0.0%]
51	[100.0%]	115	[0.5%]	179	[0.0%]	243	[0.0%]
52	[100.0%]	116	[0.0%]	180	[0.0%]	244	[0.0%]
53	[100.0%]	117	[0.0%]	181	[0.0%]	245	[0.0%]
54	[100.0%]	118	[0.0%]	182	[0.0%]	246	[0.0%]
55	[100.0%]	119	[0.0%]	183	[0.0%]	247	[0.0%]
56	[100.0%]	120	[0.0%]	184	[0.0%]	248	[0.0%]
57	[100.0%]	121	[1.4%]	185	[0.0%]	249	[0.0%]
58	[100.0%]	122	[0.0%]	186	[0.0%]	250	[0.0%]
59	[100.0%]	123	[2.3%]	187	[0.0%]	251	[0.0%]
60	[100.0%]	124	[0.5%]	188	[0.0%]	252	[0.0%]
61	[100.0%]	125	[2.7%]	189	[0.0%]	253	[0.0%]
62	[100.0%]	126	[0.0%]	190	[0.0%]	254	[0.0%]
63	[100.0%]	127	[0.0%]	191	[0.0%]	255	[0.0%]
64	[100.0%]	128	[0.0%]	192	[0.0%]	256	[0.0%]

1	[1.7%]	65	[99.4%]	129	[0.0%]	193	[0.0%]
2	[0.0%]	66	[0.0%]	130	[100.0%]	194	[0.0%]
3	[100.0%]	67	[0.0%]	131	[0.0%]	195	[0.0%]
4	[2.8%]	68	[0.0%]	132	[0.0%]	196	[100.0%]
5	[0.0%]	69	[100.0%]	133	[0.6%]	197	[0.0%]
6	[100.0%]	70	[0.0%]	134	[0.0%]	198	[0.0%]
7	[0.0%]	71	[0.0%]	135	[0.0%]	199	[100.0%]
8	[100.0%]	72	[0.0%]	136	[0.0%]	200	[0.0%]
9	[2.3%]	73	[0.0%]	137	[100.0%]	201	[0.0%]
10	[100.0%]	74	[0.0%]	138	[0.0%]	202	[0.0%]
11	[100.0%]	75	[0.0%]	139	[0.0%]	203	[0.0%]
12	[0.0%]	76	[0.6%]	140	[100.0%]	204	[0.0%]
13	[100.0%]	77	[0.0%]	141	[0.0%]	205	[0.6%]
14	[100.0%]	78	[0.0%]	142	[0.0%]	206	[0.0%]
15	[100.0%]	79	[0.0%]	143	[0.0%]	207	[0.0%]
16	[0.0%]	80	[100.0%]	144	[0.0%]	208	[0.0%]
17	[100.0%]	81	[0.0%]	145	[0.0%]	209	[0.0%]
18	[0.0%]	82	[100.0%]	146	[0.6%]	210	[0.0%]
19	[100.0%]	83	[0.0%]	147	[0.0%]	211	[0.0%]
20	[0.0%]	84	[0.0%]	148	[100.0%]	212	[0.6%]
21	[100.0%]	85	[1.1%]	149	[0.0%]	213	[0.0%]
22	[2.3%]	86	[100.0%]	150	[0.0%]	214	[0.0%]
23	[0.0%]	87	[100.0%]	151	[0.6%]	215	[0.6%]
24	[0.0%]	88	[0.0%]	152	[100.0%]	216	[0.0%]
25	[100.0%]	89	[4.5%]	153	[0.6%]	217	[0.0%]
26	[100.0%]	90	[0.0%]	154	[0.0%]	218	[0.6%]
27	[0.0%]	91	[0.6%]	155	[100.0%]	219	[0.0%]
28	[0.0%]	92	[100.0%]	156	[0.0%]	220	[0.0%]
29	[0.0%]	93	[100.0%]	157	[0.0%]	221	[0.6%]
30	[0.0%]	94	[100.0%]	158	[0.0%]	222	[0.0%]
31	[100.0%]	95	[0.0%]	159	[0.0%]	223	[0.0%]
32	[0.0%]	96	[100.0%]	160	[0.0%]	224	[0.0%]
33	[1.1%]	97	[100.0%]	161	[0.6%]	225	[0.0%]
34	[100.0%]	98	[0.0%]	162	[0.6%]	226	[0.6%]
35	[0.0%]	99	[0.0%]	163	[100.0%]	227	[0.0%]
36	[100.0%]	100	[0.0%]	164	[0.0%]	228	[0.0%]
37	[100.0%]	101	[0.0%]	165	[0.0%]	229	[0.0%]
38	[1.1%]	102	[100.0%]	166	[0.0%]	230	[0.0%]
39	[0.0%]	103	[100.0%]	167	[0.0%]	231	[0.0%]
40	[0.0%]	104	[0.0%]	168	[100.0%]	232	[0.0%]
41	[100.0%]	105	[0.0%]	169	[0.0%]	233	[0.0%]
42	[0.0%]	106	[0.0%]	170	[100.0%]	234	[0.0%]
43	[0.0%]	107	[100.0%]	171	[0.0%]	235	[0.0%]
44	[100.0%]	108	[0.6%]	172	[0.0%]	236	[0.0%]
45	[0.0%]	109	[100.0%]	173	[0.0%]	237	[0.0%]
46	[0.6%]	110	[100.0%]	174	[0.0%]	238	[0.0%]
47	[0.0%]	111	[0.0%]	175	[100.0%]	239	[0.0%]
48	[0.0%]	112	[0.0%]	176	[0.0%]	240	[100.0%]
49	[0.0%]	113	[100.0%]	177	[0.0%]	241	[0.0%]
50	[0.0%]	114	[100.0%]	178	[0.0%]	242	[0.0%]
51	[0.0%]	115	[0.0%]	179	[0.0%]	243	[100.0%]
52	[0.0%]	116	[0.0%]	180	[0.0%]	244	[100.0%]
53	[100.0%]	117	[0.0%]	181	[0.0%]	245	[0.0%]
54	[0.0%]	118	[0.6%]	182	[0.0%]	246	[100.0%]
55	[0.0%]	119	[0.0%]	183	[0.0%]	247	[100.0%]
56	[100.0%]	120	[0.0%]	184	[0.0%]	248	[0.0%]
57	[0.0%]	121	[0.0%]	185	[0.0%]	249	[100.0%]
58	[0.0%]	122	[100.0%]	186	[0.6%]	250	[0.6%]
59	[3.5%]	123	[100.0%]	187	[0.0%]	251	[0.0%]
60	[0.0%]	124	[100.0%]	188	[0.0%]	252	[0.0%]
61	[2.6%]	125	[0.6%]	189	[0.0%]	253	[100.0%]
62	[1.1%]	126	[100.0%]	190	[0.0%]	254	[0.0%]
63	[0.0%]	127	[0.6%]	191	[0.0%]	255	[0.0%]
64	[0.0%]	128	[100.0%]	192	[0.0%]	256	[0.0%]

Q&A