Wysokorozdzielcze implementacje modelu WRF Dane i optymalizacja

Prof. dr hab. inż. Mariusz J Figurski

IMGW-PIB Centrum Modelowania Meteorologicznego

Warsztaty naukowe „Modelowanie Pożarowe”, Warszawa 21-22.06.2023

Dlaczego optymalizacja jest potrzebna w modelu WRF?

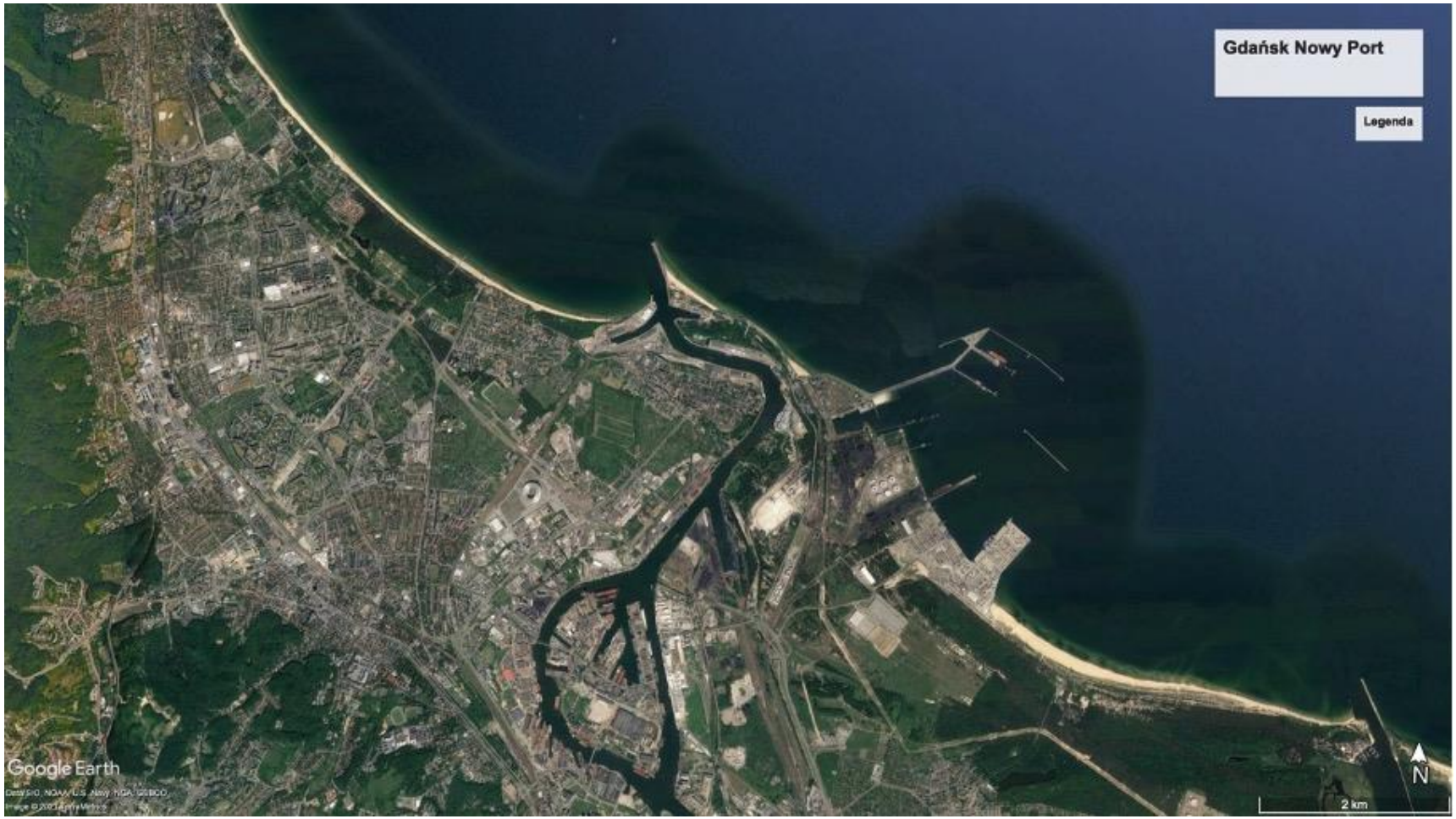
METEO
IMGW-PIB
meteo.imgw.pl

- ❑ Dopasowanie modelu do architektury superkomputera
- ❑ Duża ilość pól modelu i wysoka częstotliwość zapisywanych wyników
- ❑ Mało wydajny mechanizm zapisu i odczytu danych
- ❑ Długotrwałe symulacje np. klimatyczne
- ❑ Komunikacja międzywęzłowa
- ❑ Wysoka rozdzielczość – wielokrotne zanurzenie siatek obliczeniowych
- ❑ Problem z danymi geograficznymi dla modeli wysokiej rozdzielczości

Dane geograficzne w modelu WRF

METEO
IMGW-PIB
meteo.imgw.pl

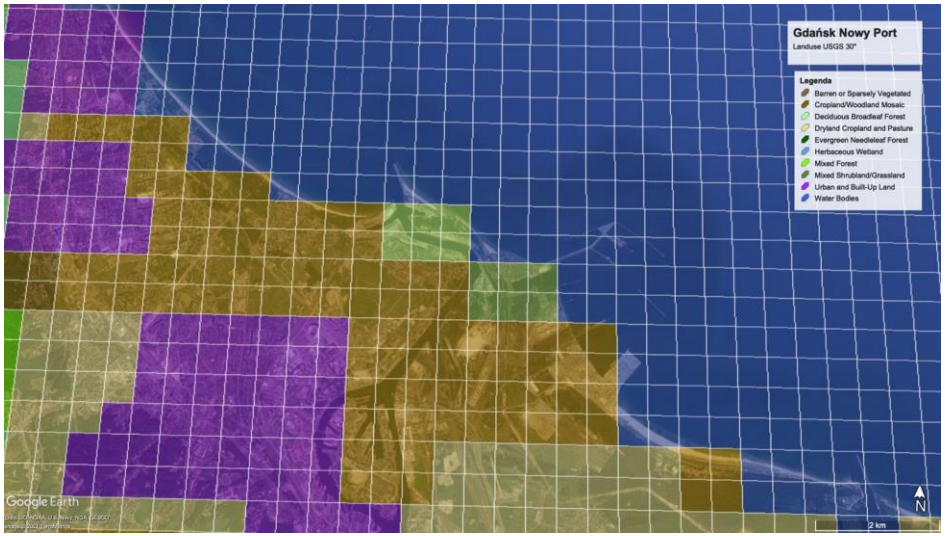
Ogólna zasada adaptacji danych geograficznych w modelu WRF



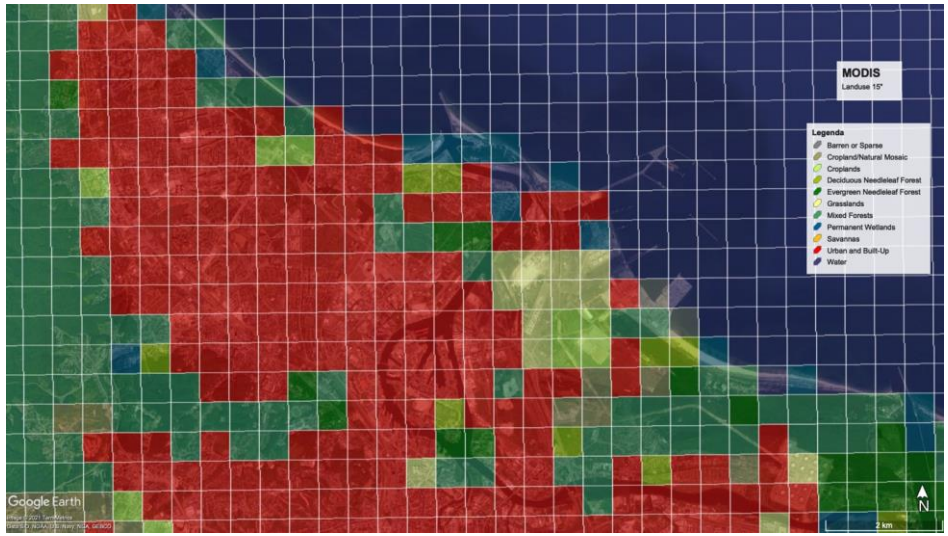
Dane geograficzne w modelu WRF

METEO IMGW-PIB meteo.imgw.pl

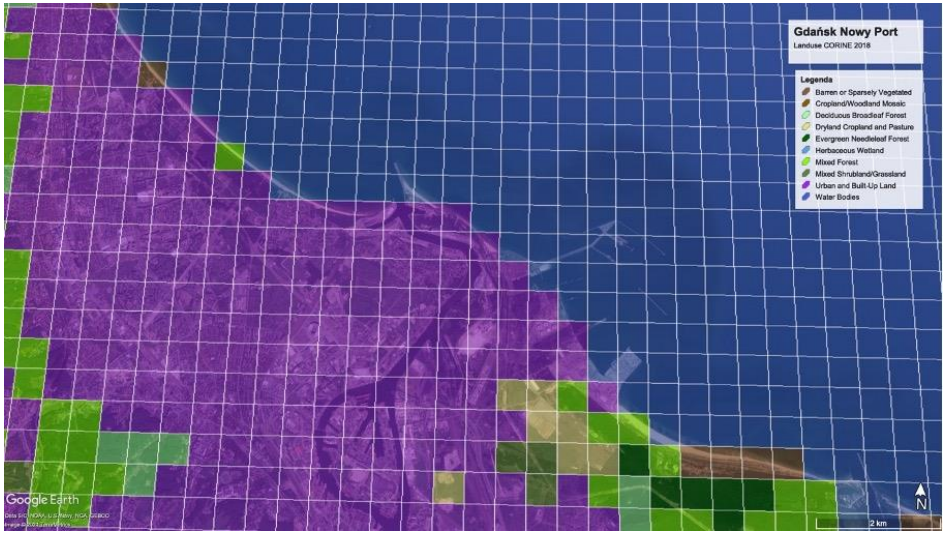
Aktualność i wiarygodność danych – pokrycie terenu



USGS 30”



MODIS 15”



CORINE 2018 100m

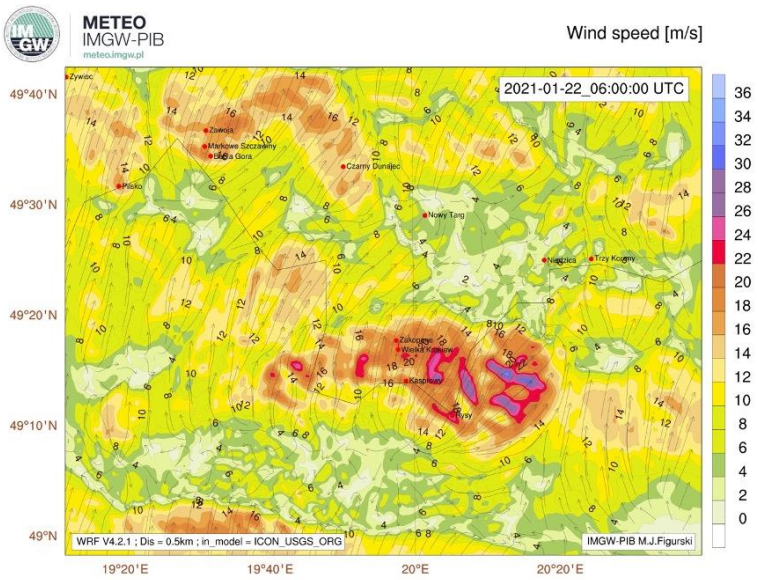
Problemy

- Reklasyfikacja typów do MODIS lub USGS
- Dostosowanie do układu WGS84
- Ograniczenie do państw EU

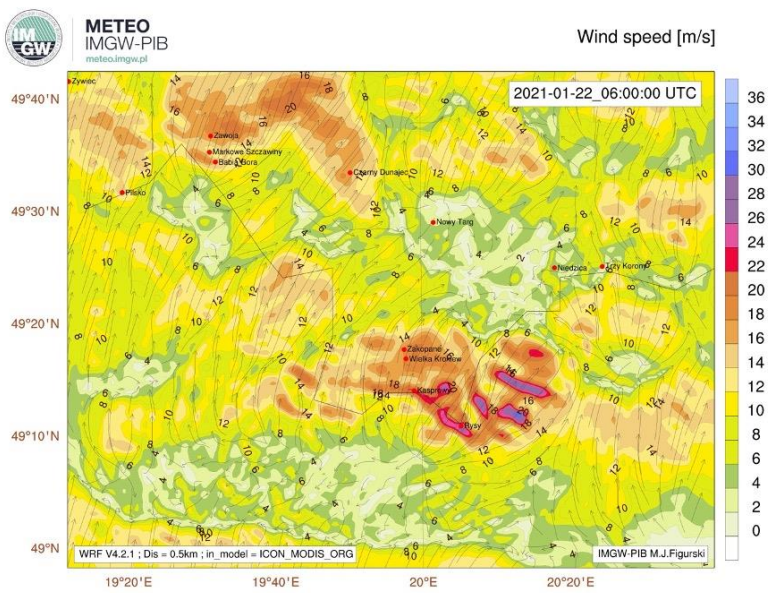
Zalety

- Zweryfikowane dane
- Aktualne dane, szczególnie aglomeracje

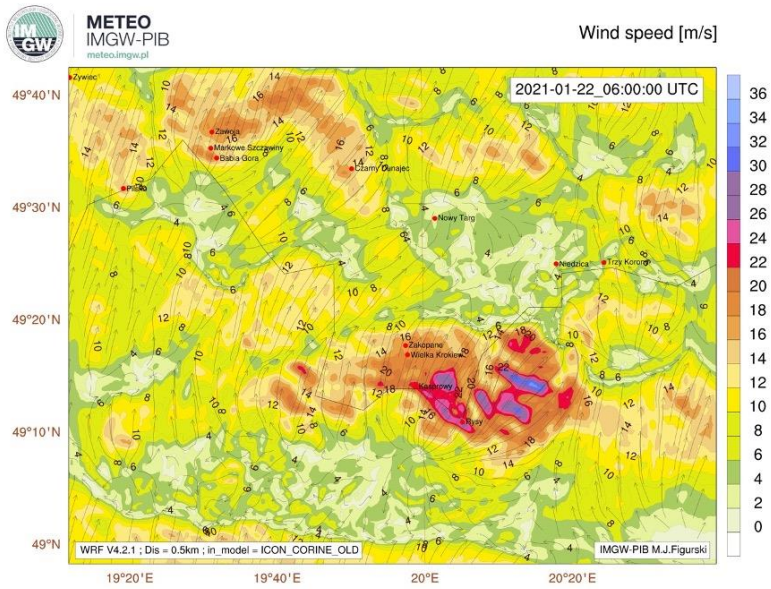
Dane geograficzne w modelu WRF



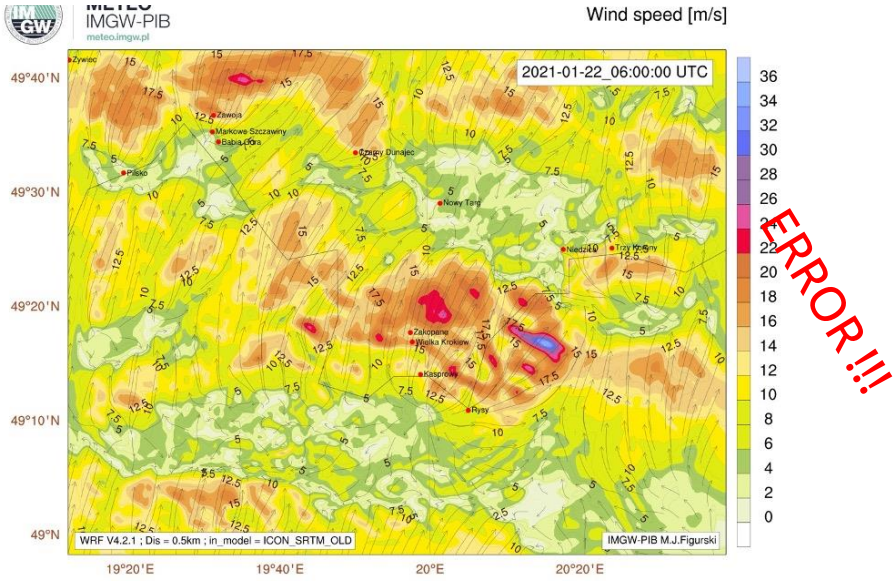
USGS 30″ + SRTM 1″



MODIS 15″ + SRTM 1″



CORINE 2018 100m + SRTM 1″



MODIS 15″ + SRTM 1″

Intel Hyper-threading

- Niektóre aplikacje HPC uzyskują lepszą wydajność poprzez wyłączenie jednoczesnej wielowątkowości (SMT).
- Jednoczesna wielowątkowość, powszechnie znana jako Intel Hyper-threading , przydziela dwa wirtualne rdzenie (vCPU) na rdzeń fizyczny w węźle.
- W przypadku wielu ogólnych zadań obliczeniowych lub zadań wymagających dużej ilości operacji we/wy, SMT może znacznie zwiększyć przepustowość aplikacji.
- W przypadku zadań związanych z obliczeniami, w których oba rdzenie wirtualne są powiązane z obliczeniami, SMT może obniżyć ogólną wydajność aplikacji i dodać nieprzewidywalne odchylenia od zadań.
- Wyłączenie SMT dla numerycznych modeli pogody zapewnia bardziej przewidywalną wydajność i może skrócić czas pracy.

Dostosowanie limitów zasobów systemowych

Systemy uniksowe mają domyślne limity zasobów systemowych, takich jak otwarte pliki i liczba procesów, z których może korzystać każdy użytkownik. Limity te uniemożliwiają jednemu użytkownikowi monopolizowanie zasobów systemu i wpływanie na pracę innych użytkowników. Jednak w kontekście HPC te limity są zwykle niepotrzebne, ponieważ węzły obliczeniowe w klastrze nie są bezpośrednio udostępniane przez użytkowników.

Można dostosować limity użytkowników, edytując plik /etc/security/limits.conf i ponownie logując się do węzła.

Dostosowując limity użytkowników, zmień wartości następujących limitów:

- nproc- maksymalna liczba procesów
- memlock- maksymalna przestrzeń adresowa zamknięta w pamięci (KB)
- stack- maksymalny rozmiar stosu (KB)
- nofile- maksymalna liczba otwartych plików
- cpu- maksymalny czas procesora (minuty)
- rtprio- maksymalny priorytet w czasie rzeczywistym dozwolony dla procesów nieuprzywilejowanych (Linux 2.6.12 i nowsze)

OpenMPI – problemy na superkomputerze Tryton

- Unified Communication X (UCX) to struktura interfejsów API komunikacji dla HPC. Jest zoptymalizowana pod kątem komunikacji MPI za pośrednictwem rozwiązania InfiniBand i współpracuje z wieloma implementacjami MPI, takimi jak OpenMPI i MPICH
- Obecnie dla superkomputrera Tryton optymalnym rozwiązaniem dla modelu WRF jest użycie wersji OpenMPI 1.10.7 lub Intel MPI. W przypadku Intel MPI problem z poprawną konfiguracją modelu WRF.
- W wersji OpenMPI 4.x występują błędy związane z kompatybilnością biblioteki InfiniBand Mellanox. Konfiguracja z biblioteką UCX na Trytonie powoduje błędy wykonania modelu:

  WARNING: There was an error initializing an OpenFabrics device.
- Domyślnie dla Open MPI 4.0 i nowszych, porty infiniband na urządzeniu nie są domyślnie używane. Intencją jest użycie UCX dla tych urządzeń. Tę strategię można zastąpić, ustawiając parametr MCA btl_openib_allow_ib na wartość true.
- Porównanie wydajności modelu WRF z OpenMPI v. 1.10.7 i v. 4.x pokazuje, że nowsza wersja działa około 10% wolniej.
- **W przetwarzaniu równoległym procesy zawsze czekają na najwolniejsze zadanie. Co to powoduje w rzeczywistości modelu?**

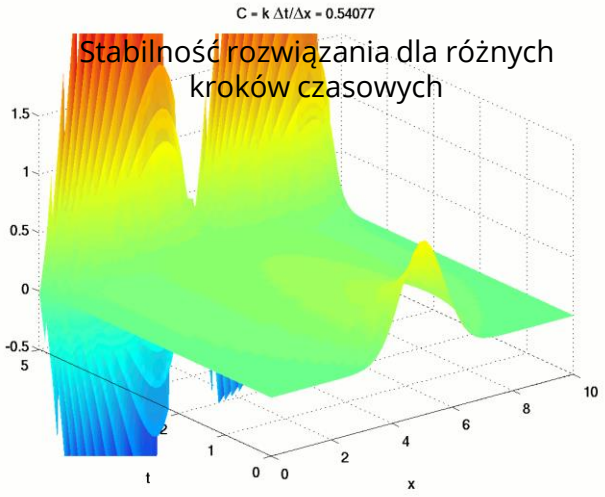
Warunek Couranta

- Warunek Couranta-Friedrichsa-Lewy'ego (CFL) matematyczny warunek zbieżności numerycznych metod rozwiązywania równań różniczkowych cząstkowych. Zwykle dotyczy równań różniczkowych z członem adwekcyjnym opisującym propagację fal np. równania adwekcji. W praktyce wpływa na wybór długości kroku czasowego modelu.

- Fizyczne znaczenie warunku CFL można przedstawić następująco. Jeżeli oryginalne równanie różniczkowe opisuje propagację fali, to w modelu numerycznym, w którym przestrzeń ciągłą przybliżono dyskretną siatką punktów, fala musi przechodzić pomiędzy sąsiednimi punktami siatki w czasie nie dłuższym niż czas potrzebny rzeczywistej fali na pokonanie tej samej odległości:

$$\Delta T < \frac{\Delta x}{U_{max}}$$

ΔT – *numeryczny krok czasowy*,
Δx – *warość stałej w modelu*,
U_{max} – *maksymalna prędkość fali*

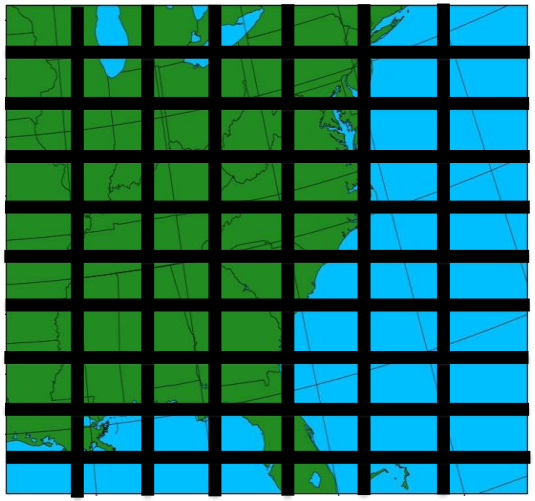


Wraz ze zmniejszaniem odległości między punktami siatki obliczeniowej redukcji ulega też maksymalna wartość kroku czasowego używanego w symulacji

Dekompozycja siatek obliczeniowych

METEO IMGW-PIB meteo.imgw.pl

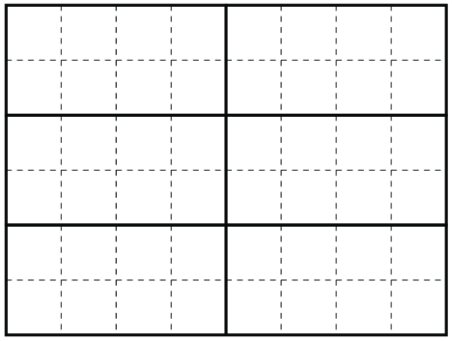
- 70 zadań
- 10 (j) x 7 (i)



- 71 tasks
- 71 (j) x 1 (i)

WRF został napisany z myślą o kompilacji MPI, OpenMP i OpenMP-MPI, co ma wpływ na sposób, w jaki rozkłada domenę modelu

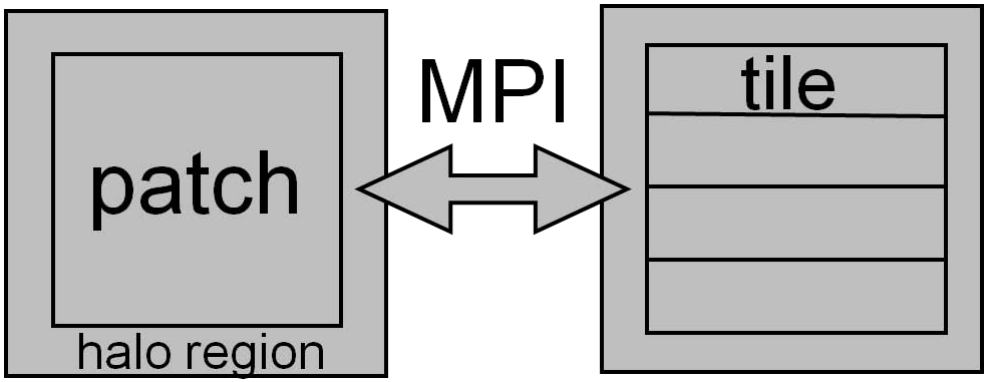
- WRF domyślnie wyznacza wartości nproc_x i nproc_y z pierwiastka kwadratowego przydzielonej ilość rdzeni/procesorów.
- Jeśli jest to niemożliwe używane są wartości zbliżone do siebie.
- Dekompozycja odnosi się do siatki nadrzędnej 2D.

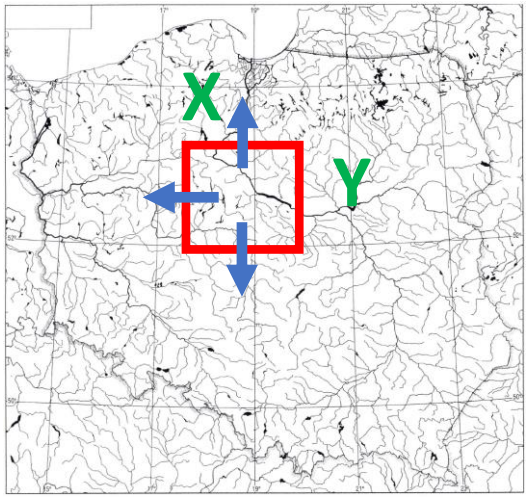


Ilustracja rozkładu domeny w ramach WRF odpowiednia dla zadania w trybie mieszanym z sześcioma procesami MPI, każdy z ośmioma wątkami OpenMP. Łaty są rysowane pogrubionymi, ciągłymi liniami, a płytki z liniami przerywanymi.

- Wszystkie domeny w modelu WRF używają dokładnie takiej samej liczby procesorów, z taką samą dekompozycją procesorów.
- Wybór domen z grubą siatką i domen z drobną siatką, które mają dużą różnicę w ilości komórek siatki, negatywnie wpływa na wydajność taktowania
- Mała siatka zgrubna (w porównaniu z zanurzoną) ogranicza liczbę procesorów, które mogą być wykorzystane przez domenę z siatką drobną (droga domena)
- Drobna siatka jest kosztowną domeną

Dekompozycja w modelu WRF - komunikacja





Domena obliczeniowa jest podzielona na rozłączne prostokątne łaty. Każda łata (patch) jest aktualizowana przez pojedynczy proces MPI (równoległość pamięci rozproszonej), a proces może odczytywać dane archiwalne w postaci paska wokół łaty, zwanego regionem halo. Komunikacja między łatami odbywa się za pomocą wywołań halo do infrastruktury równoległej RSL (Parallel Runtime System Library For Regional Atmospheric Models With Nesting) (Michalakes, 2000), które aktualizują regiony halo o wartości pochodzące z sąsiednich łat. Każda łata może być podzielona na kafelki, które są wykonywane w osobnych wątkach OpenMP wątki (równoległość pamięci współdzielonej). Zgodnie z konwencją kodowania WRF (Grupa Robocza WRF 2, 2007), jądra obliczeniowe działają w pojedynczym kafelku. Mogą one odczytywać wartości tablicowe z paska poza granicami kafla, ale nie jest dozwolona żadna jawna komunikacja. Tablice trójwymiarowe podzielone są na łaty i kafelki w płaszczyźnie poziomej

Ilu procesorów użyć do modelowania?

Kiedy rozkładamy domenę na mniejsze części, rozdzielamy obliczenia na dodatkowe rdzenie/procesory

```
mpirun -np ??? wrf.exe
```

Aby wybrać odpowiednią liczbę procesorów, należy wziąć pod uwagę dekompozycję procesów w zależności od wielkości domen, szczególnie dla wielokrotnego zagnieżdżenia.

Pytania:

- W jaki sposób ustalamy, ile procesorów należy wykorzystać do pracy?
- Jaki jest kształt powstałych rozłożonych siatek?
- Jakie są konsekwencje wydajnościowe tych decyzji?

Ilu procesorów użyć do modelowania?

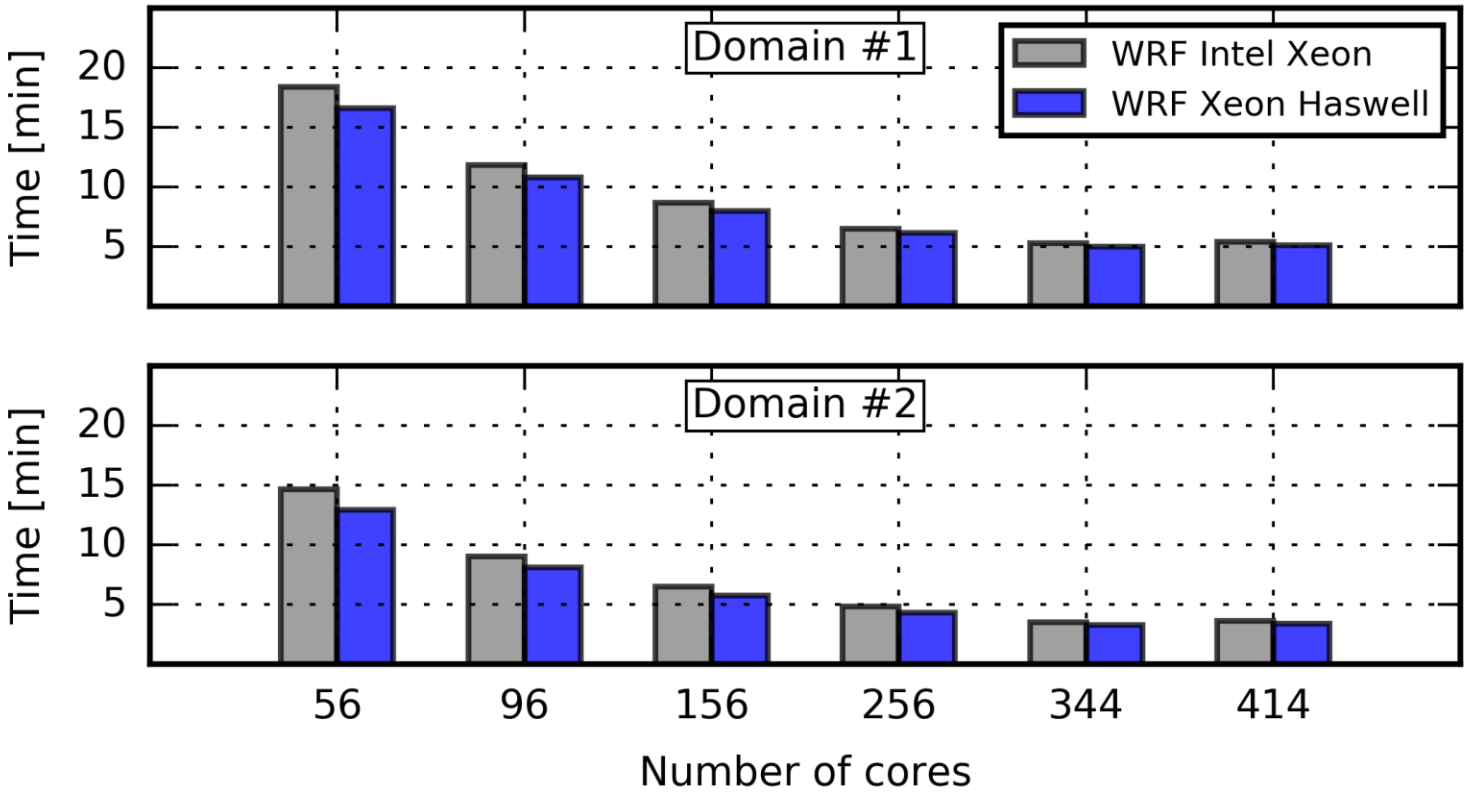
Jakiej ilości procesorów użyć do obliczeń?

- Minimalna liczba procesorów jest zależna od ilości dostępnej pamięci !
- Dla 1500x1500 węzłów modelu wymagane są co najmniej 4 węzły (4x24 rdzeni procesora), ale tylko 2 węzły były wymagane podczas korzystania z węzłów o dużej ilości pamięci (kilkaset GB).
- Model WRF (z przykładu) nie działałby z jednym węzłem, wymagałby zbyt dużej ilości pamięci
- Maksymalna ilość rdzeni/procesorów zależy od komunikacji wewnątrz modelu WRF
- Model zatrzymuje się jeśli na jeden wątek przypada nie więcej niż 10 komórek siatki obliczeniowej w każdym kierunku (i,j) (wschód-zachód, południe-północ)
- W przypadku 1500x1500 węzłów modelu możemy mieć w każdym kierunku (i,j) 150 jednostek obliczeniowych (łat) o szerokości 10 komórek siatki
- Z tego wynika, że maksymalnie można użyć 150x150 = 22500 rdzeni procesora
- Zlecana liczba procesorów jest szacowana z wydajności czasowej, jaką model WRF może zapewnić z terminowością wykonania rozwiązania/zadania
- Zazwyczaj mniejsza liczba procesorów pozwala bardziej efektywnie wykorzystać superkomputer (skalowalność oraz operacje we/wy)
- Dopóki jest wystarczająca liczba procesorów (dla pamięci) i nie jest ich zbyt dużo na zdefiniowaną domenę, to rozwiązanie modelu WRF jest poprawne.

Zalecenia praktyczne

- Aby wybrać odpowiednią liczbę procesorów, należy wziąć pod uwagę dekompozycję procesów w zależności od wielkości domen. Płytki (łatki) powinny być kształtem być zbliżone do kwadratów, ale można od tego nieco odejść.
- Nie można doprowadzić, aby cała płytka(łatki) była regionami halo, ponieważ będzie potrzebna rzeczywista przestrzeń do obliczeń pośrodku każdej płytki (łatki). Jeśli przestrzeń obliczeniowa nie istnieje, może to spowodować awarię modelu lub nierealistyczne wyniki.
- Największa liczba procesorów, których powinno się użyć, powinna opierać się na najmniejszej domenie, a najmniejsza liczba procesorów powinna opierać się na największej domenie. Dlatego ważne jest, aby nie mieć domen, które różnią się zbytnio rozmiarem (wielkością przestrzenną siatki). Nie chcemy używać zbyt małej liczby procesorów, ponieważ może to spowodować bardzo wolne działanie (lub niemożliwe - model może się zawiesić), więc należy to również wziąć pod uwagę.
- Do określenia optymalnej liczby rdzenie/procesorów można użyć przybliżonych równań uzyskanych empirycznie:
  - Dla domeny o najmniejszym rozmiarze:
    ((e_we)/25) * ((e_sn)/25) = największa liczba procesorów, których można użyć
  - Dla domeny o największym rozmiarze:
    ((e_we)/100) * ((e_sn)/100) = najmniejsza liczba procesorów, których można użyć
  - Dla domeny modelu 1500x1500 węzłów największa i najmniejsza liczba rdzeni procesorów wynoszą odpowiednio 3600 i 225.

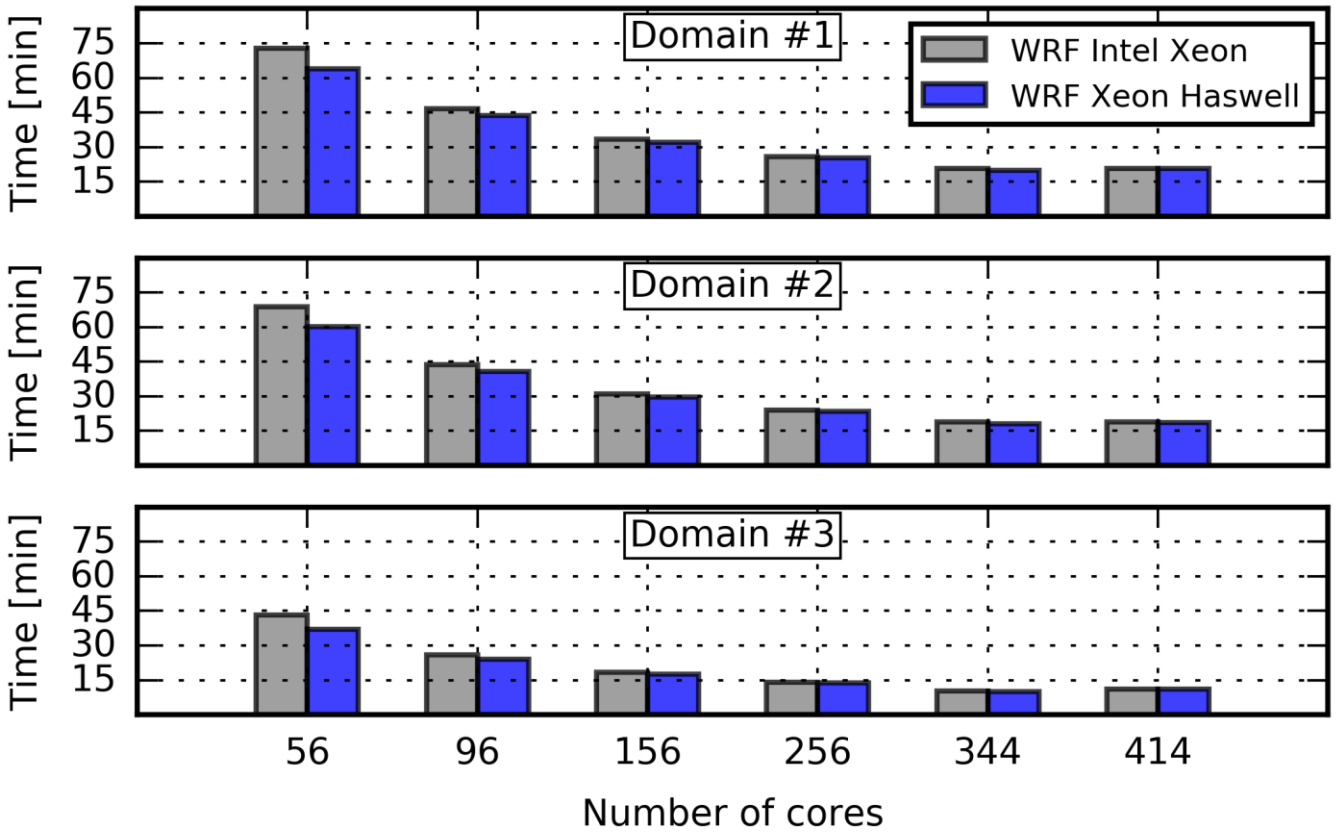
Skalowalność procesów (WRF METEOPG)



Dwie domeny obliczeniowe

- Dobra skalowalność procesów do około 200 rdzeni
- Optymalizacja kodu dla procesorów Haswell
- Przyspieszenie obliczeń o koło 12%
- Problemy z niektórymi modelami mikrofizyki

Skalowalność procesów (WRF METEOPG)



Trzy domeny obliczeniowe

- Optymalizacja dla procesorów Haswell
- Dla 3 siatek optymalizacja modelu dla procesorów Haswell przyspiesza obliczenia o około 20%. Skalowalność procesów do 350 rdzeni.
- Przyspieszenie jest obserwowane tylko dla procesów dobrze skalowalnych.

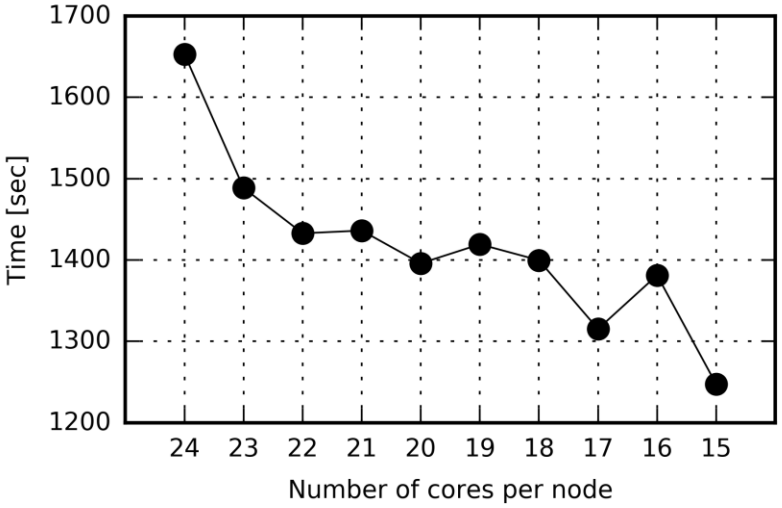
Powinowactwo procesorów

METEO IMGW-PIB meteo.imgw.pl

Powinowactwo procesorowe (znane również jako przypinanie procesora) to proces przypisywania działających programów do jednego wątku (wirtualnego rdzenia), zamiast pozwalania mu na działanie ze wszystkimi wątkami CPU. Ustawienie powinowactwa procesu jest korzystne, ponieważ pozwoli użytkownikom dokładnie określić, ile zasobów zużywa program.

Zdiagnozowany problem

Wykorzystanie wszystkich rdzeni procesora spowalnia procesy obliczeniowe 11% !!!

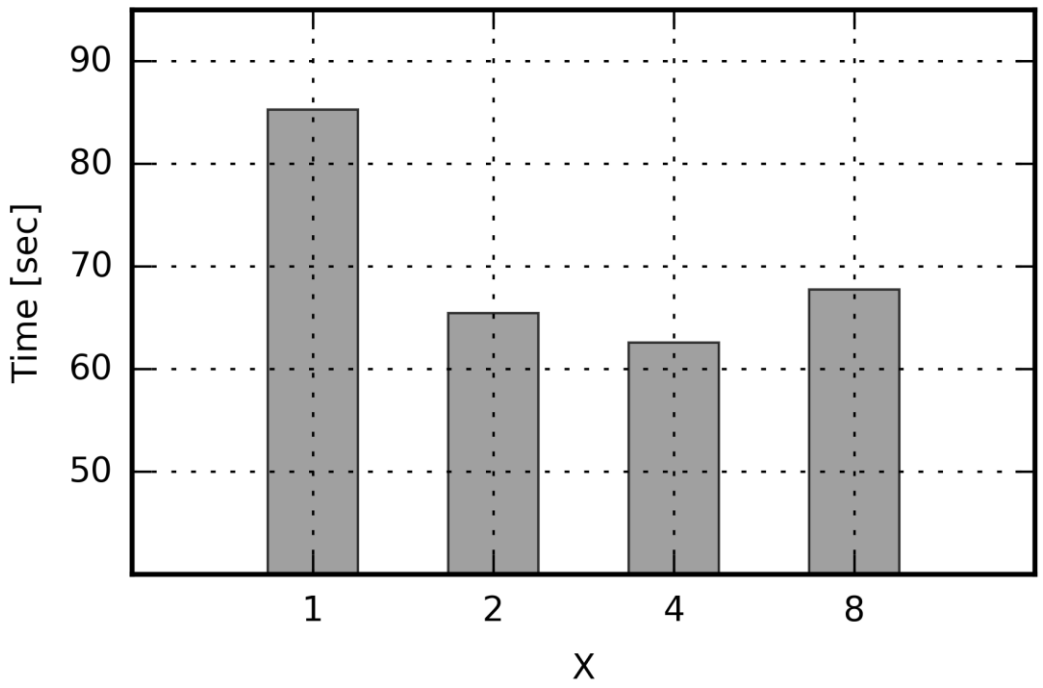


- Rozwiązanie1.
- Optymalne rozwiązanie, zmniejszenie z 24 do 22-18 rdzeni/procesor.
- Rozwiązanie2.
- Zmiana przyporządkowanego rdzeni procesorów do konkretnego procesu lub wątku, działa poprawnie jeśli jest wystarczająca ilość pamięci podręcznej.

mpirun –bycore –bind-to-core –report-bindings ./wrf.exe

Procedura powyższa nie zezwala na migrację procesów. Jest to ważne ponieważ jeśli proces migruje, nie może odnaleźć wcześniej zapisanych danych w nowej pamięci podręcznej.

Dekompozycja w modelu WRF - komunikacja



- Z praktycznego punku widzenia przyjęcie takich samych wartości nproc_x i nproc_y nie jest optymalne. WRF lepiej pracuje jeśli dekompozycja ma bardziej prostokątny charakter. Prowadzi to do lepszego wykorzystania pamięci podręcznej i bardziej efektywnej komunikacji.
- W analizowanym przypadku domyślna wartość nproc_y=8 (nproc_x=8), zmieniając tą wartość na 16 (nproc_y=4), czas obliczeń zmniejsza się o około 8%.

Problem kroku czasowego w modelu WRF

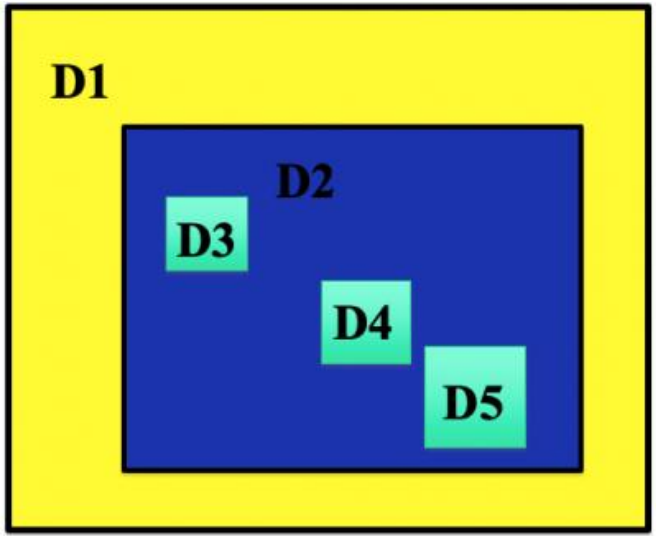
METEO IMGW-PIB meteo.imgw.pl

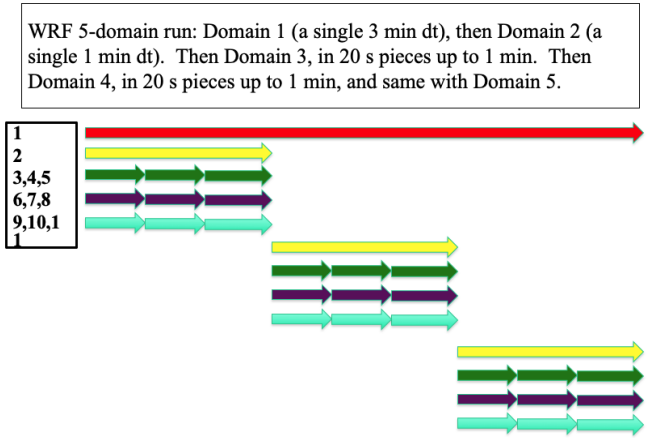
- **Statyczny** zalecane użycie zależności DX*6; DX – rozmiar oczka siatki modelu w km.
- **Adaptacyjny** krok czasowy to metoda maksymalizacji kroku czasowego, z którego może korzystać model, przy jednoczesnym utrzymaniu stabilności numerycznej modelu. Krok czasowy modelu jest dostosowywany w oparciu o kryterium stabilności poziomej i pionowej w całej domenie (warunek Couranta-Friedrichsa-Lewy'ego (CFL)). Poniższy zestaw wartości zazwyczaj działa dobrze.

`use_adaptive_time_step` = .true.
`step_to_output_time` = .true. (ale domeny zagnieżdżone mogą nadal zapisywać dane wyjściowe w żądanym czasie. Aby to zrekompensować, należy użyć *adjust_output_times* = *.true* .)
`target_cfl` = 0.85, 0.85, 0.85, sterowanie poziomym krokiem dla warunku CFL
`target_hcfl` = 0.65, 0.65, 0.65, sterowanie pionowym krokiem dla warunku CFL
`max_step_increase_pct` = 5, 51, 51 (duża wartość procentowa dla gniazda pozwala na większą swobodę regulacji kroku czasu dla gniazda)
`starting_time_step` = 38, 15, 5, dowolny krok czasowy, który uzyje model w czasie startu, 4*DX
`max_time_step` : maksymalny krok czasowy, domyślna wartość 8*DX
`min_time_step` : minimalny krok czasowy, domyślna wartość 3*DX
`adaptation_domain` : która domena steruje adaptacyjnym krokiem czasu (liczba całkowita)

Jeśli pojawiają się artefakty numeryczne:
Błędnie dobrane kroki czasowe
&domains
perturb_input = .true.

Szacowanie wydajności obliczeniowej





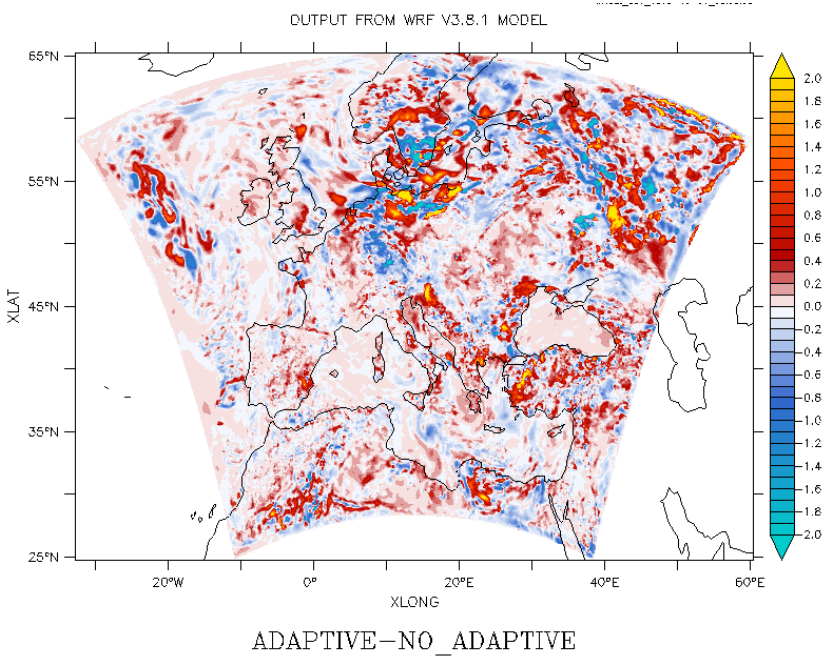
1. Domena 1 wykonuje pojedynczy krok czasowy przez 3 minuty (czerwona strzałka).
2. Domena 2 wykonuje pojedynczy krok czasowy przez 1 minutę (żółta strzałka). Zanim domena 2 będzie mogła wykonać drugi krok czasowy, wszystkie elementy podrzędne domeny 2 muszą nadrobić zaległości (ukończyć obliczenia w danym kroku).
3. Domena 3 wymaga kilku kroków czasowych, aby dogonić domenę 2 (20-sekundowy krok czasowy).
4. Domena 4 wymaga kilka kroków czasowych, aby dogonić domenę 2.
5. Domeny 5 wymaga kilka kroków czasowych, aby dogonić domenę 2.
6. Domeny 3, 4 i 5 przekazują informację zwrotną do domeny 2.
7. Kroki 2–6 są powtarzane, aż domena 2 dogoni domenę 1.
8. Domena 2 przekazuje informację zwrotną do domeny 1.
9. Kroki 1 - 8 są powtarzane, aż wszystkie domeny zakończą całą symulację.

Stosunek odległości między siatkami oraz stosunek kroków czasowych wynoszą 3:1. Załóżmy, że liczba komórek siatki jest identyczna w każdej z domen (np. wszystkie mają wymiary 100x100 lub 500x500, nie ma to znaczenia). Koszt działania domeny 1 to koszt jednostkowy.

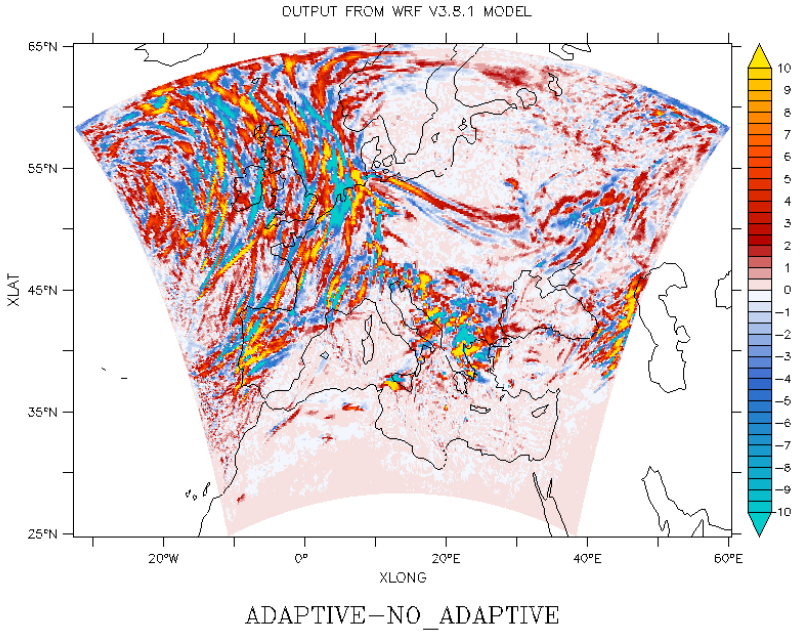
Bez narzutu na zagnieżdżanie można by oczekiwać, że przebieg w pięciu domenach będzie 31x dłuższy niż w przypadku pojedynczej domeny (1 + 3 + 9 + 9 + 9).

Problem kroku czasowego w modelu WRF

Statyczny, czy adaptacyjny ?



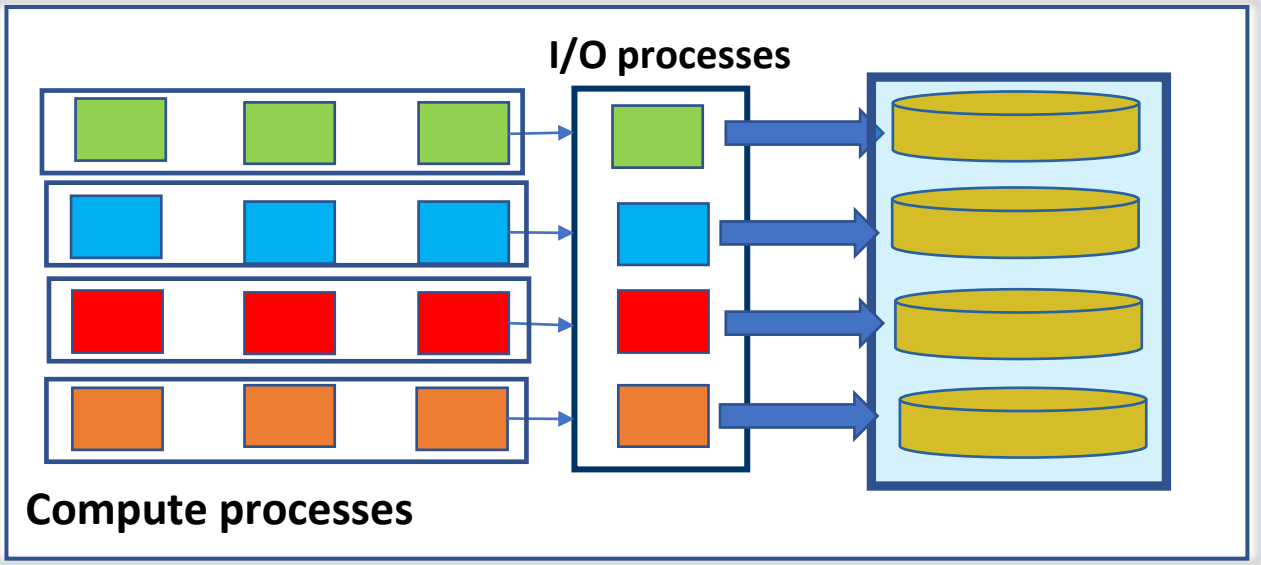
Tempertaura 2m. Różnica między symulacją z krokiem adaptacyjnym i stałą wartością dt. Po 6 dniach prognozy.



Różnica w opadzie skumulowanym między symulacją z wykorzystaniem adaptacyjnego kroku i stałą wartością dt. Po 9 dniach prognozy

Optymalizacja we/wy modelu WRF

Domyślny zapis synchroniczny – wada zatrzymanie wątków do momentu dotarcia danych



Rozwiązanie 1 – synchroniczny dla każdego wątku

- Wykorzystanie pnetcdf, każdy proces MPI zapisuje własny plik. Poprawa zależy od wielkości danych na proces. Dla modelu METEOPG wydajność zapisu danych wzrosła 10 razy dla 528 rdzeni.

Rozwiązanie 2 – zapis asynchroniczny

- Rozwiązanie: Połączenie pnetcdf z warstwą Tryton MPI oznacza, że szeregi MPI są łączone w grupy, a następnie jeden z każdej grupy agregator wykonuje zapis do pliku.

nproc = nproc_x*nproc_y+(nio_groups*nio_task_per_group),

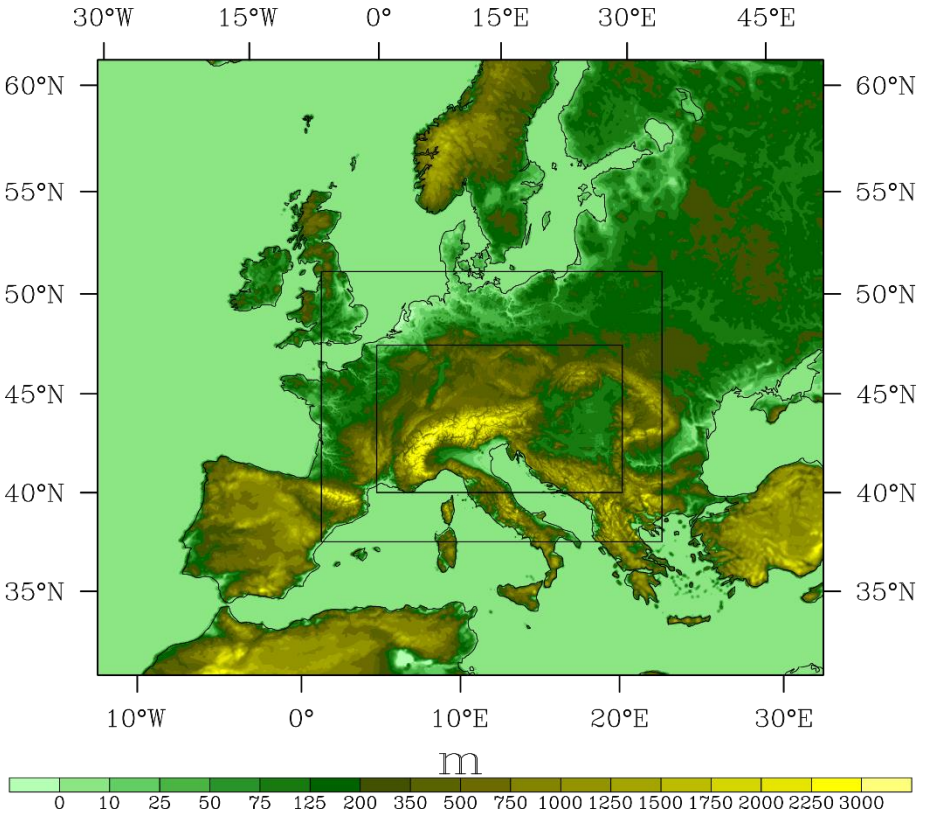
- wartość nio_task_per_group nie może przekroczyć wartości nproc_y. Optymalne rozwiązanie nproc_y powinno być wielokrotnością nio_task_per_group. Wada dodatkowe procesowy do obsługi we/wy

- Wada: musimy połączyć wszystkie pliki. WRF oficjalnie nie wspiera tego rozwiązania. Powstaje tyle plików ile jest użytych rdzeni.

Benchmark superkomputerów

- **Arctic Region Supercomputing Center WRF Benchmarking.**
  Dane dla siatek zanurzonych obejmujące obszar Europy i Alaski. WRF 3.2.1.
  Symulacja o długości 3h na bazie danych z 2007 roku.
- **Dane obecnie niedostępne.**

Rozdzielczość	Liczba węzłów
7.2 km	585x495x63 = 18.2 mln
2.4 km	823x652x63 = 33.8 mln
800 m	1777x1066x63 = 119.3 mln



Benchmark superkomputerów

METEO IMGW-PIB meteo.imgw.pl

Testowane platformy superkomputerowe

VSC	PACMAN	CHUGACH	KRAKEN
HPC academic system University of Vienna Sun Fire X2270 compute nodes, each equipped with 2 Quadcore processors (Intel, X5550, 2.66 GHz) and 24 GB memory (3 GB per core). Infiniband QDR network (40 Gbps). Filesystem ext3	**Academic system , Arctic Region Supercomputing Center. Pacific Area Climate Monitoring and Analysis Network (PACMAN)**. Sixteen-core compute nodes consisting of 2 eight-core 2.3 GHz AMD Opteron processors with 64 GB memory (4 GB per core). Mellanox QDR Infiniband interconnect.	**Cray XE6 currently administered by ARSC for the DoD High Performance Computing and Modernization Program.** 16-core compute nodes consisting of 2 eight-core 2.3 GHz AMD Opteron processors with 32 GB memory (2 GB per core). Cray Gemini interconnect. Lustre scalable filesystem used on compute nodes.	**Cray XT5 at National Institute for Computational Sciences.** 12-core compute nodes consisting of 2 six-core 2.6 GHz AMD Opteron processors with 16 GB memory (1.5 GB per core). Cray SeaStar2+ interconnect. Lustre filesystem used on compute nodes.
			
Austria, Wiedeń	Fairbanks, USA	USA	Tennessee, USA

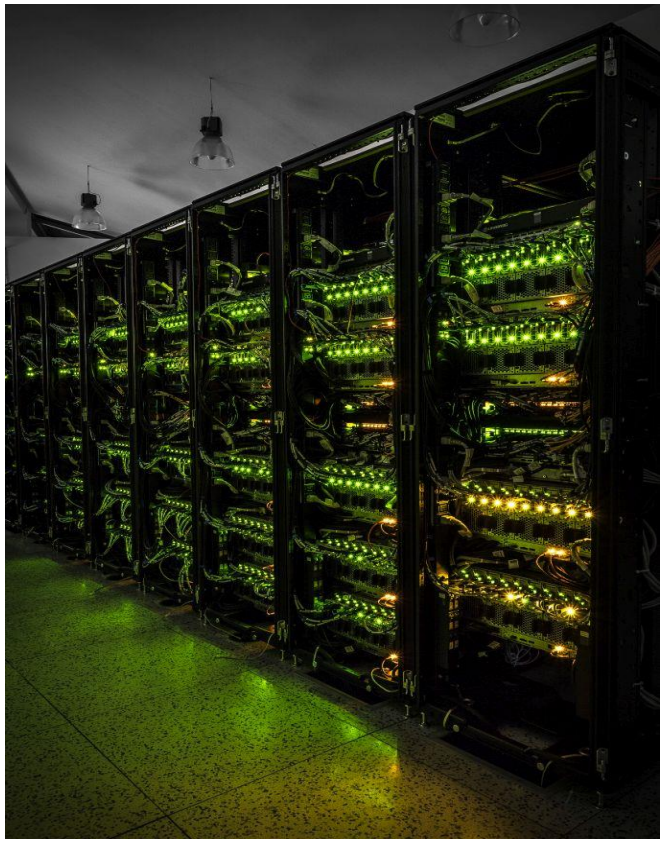
Benchmark superkomputerów

METEO IMGW-PIB meteo.imgw.pl

Testowane platformy superkomputerowe

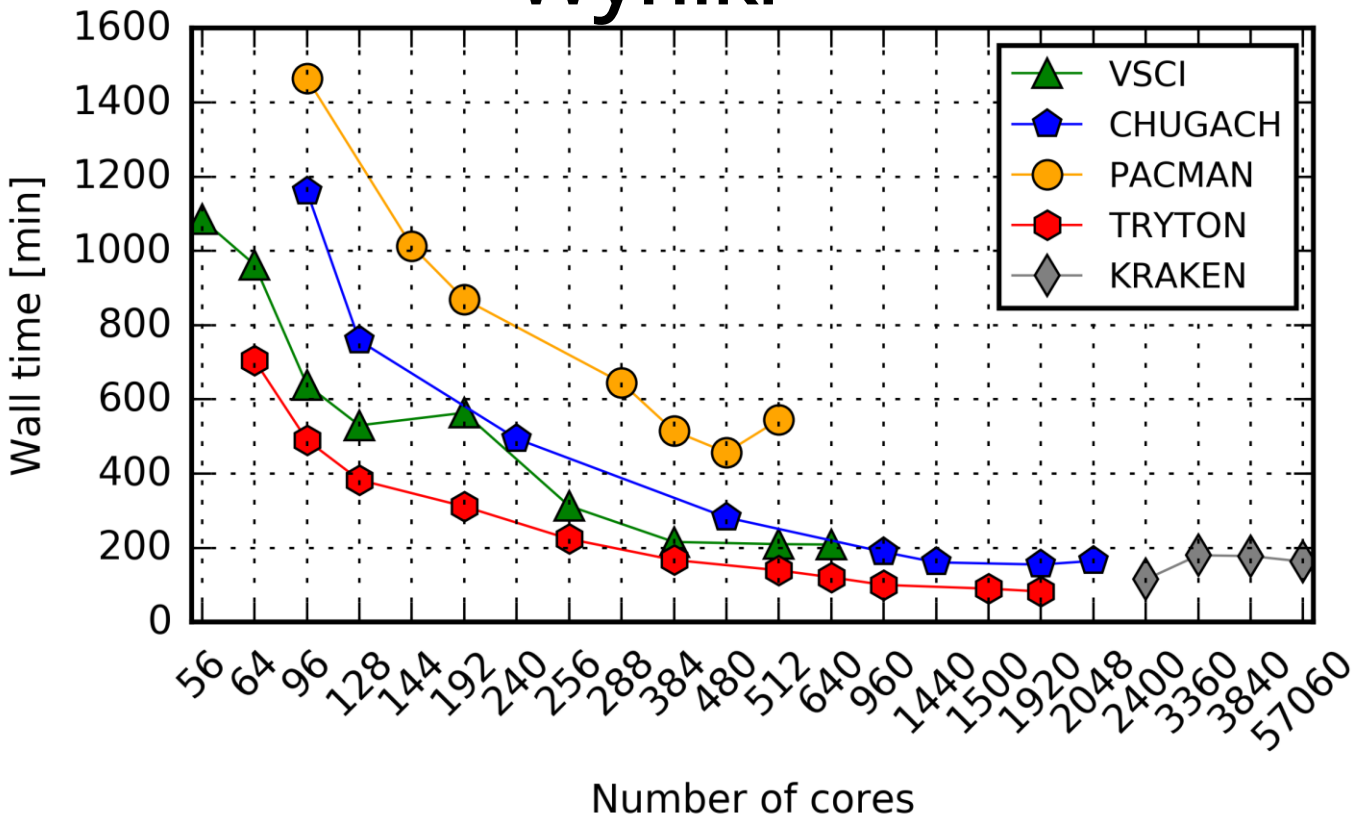
Tryton - superkomputer w Centrum Informatycznym Trójmiejskiej Akademickiej Sieci Komputerowej (CI TASK) o architekturze klastrowej o następujących parametrach:

Procesory:	Intel® Xeon® Processor E5 v3 @ 2,3 GHz, 12-core (Haswell), 3MB cache
Akceleratory:	Nvidia Tesla, Intel Xeon Phi, AMD FirePro
Pamięć:	128/256 GB RAM DDR4 na serwer
Sieć:	InfiniBand FDR 56 Gb/s, topologia fat tree, przełączniki Mellanox
Razem:	1607 serwerów, 3214 procesorów, 38568 rdzeni, 48 akceleratorów, 218 TB RAM
Szafy:	40 szt.
System operacyjny:	linux
Moc obliczeniowa:	1,48 PFLOPS



Benchmark superkomputerów

Wyniki



- Wykorzystanie pamięci operacyjnej: ponad 1.5TB.
- Symulacje wykonane modelem WRF 3.8.1 oraz 3.2.1. Zwiększenie ilości rdzeni z 512 do 1920 nie przyspiesza symulacji – problem skalowalności procesów.
- Brak danych dla nowszych wersji modelu WRF np. 4.x

„Nie chcemy nikomu modelować życia, ale chcemy coraz lepiej modelować pogodę, aby żyło się lepiej”

Prof. dr hab. inż. Mariusz J Figurski

Laboratorium Zaawansowanych Metod Modelowania Meteorologicznego IMGW-PIB